Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени первого Президента России Б. Н. Ельцина»

УТВЕРЖДАЮ

Директор по образовательной деятельности

_____ С.Т. Князев

«24» декабря 2021 г.

Математические основы искусственного интеллекта

Учебно-методические материалы по направлению подготовки
09.04.01 Информатика и вычислительная техника
Образовательная программа «Инженерия искусственного интеллекта»

Екатеринбург

2021

РАЗРАБОТЧИКИ УЧЕБНО-МЕТОДИЧЕСКИХ МАТЕРИАЛОВ

Доцент, кандидат физико-математических наук		Солодушкин Святослав Игоревич
Доцент, кандидат физико-математических наук		Юманова Ирина Фарисовна

Содержание

Предисловие

В настоящее время в связи с цифровизацией многих сфер деятельности поток информации, доступной исследователям, стал поистине огромным. При этом справедливым остается тезис: данных много, а знаний мало. В связи с этим уделяется большое внимание извлечению знаний из неструктурированных, зашумленных первичных данных.

Мы свидетели становления, по сути, нового направления в науке — анализа данных. Соответственно, бизнес и наука ставят перед высшим образованием задачу подготовки специалистов, способных этот анализ данных проводить. Наряду с нейронными сетями важным методом анализа данных является статистика.

Руководство по проведению лекций «Математические основы искусственного интеллекта» написано авторами на основе опыта чтения соответствующих курсов в Уральском федеральной университете. Цель курса — изучение методов сбора и первичной обработки информации, проверки статистических гипотез, анализа статистических связей.

Пособие разбито на главы. Каждая глава соответствует одной рассматриваемой на занятиях теме и содержит необходимые теоретические сведения, примеры, всесторонне иллюстрирующие теорию. В конце глав даются задания для самоконтроля.

Авторы пособия — математики по образованию — в течение многих лет участвовали в клинических исследованиях, проводили статистический анализ медицинских данных. Большинство примеров, представленных в пособии, являются реальными и взяты авторами из собственной практики.

1. Выборочный метод

1.1. Предмет и задачи статистики

Прикладная статистика — раздел математики, в котором разрабатываются методы регистрации, описания и анализа данных наблюдений и экспериментов с целью построения вероятностных моделей массовых случайных явлений.

Предметом прикладной статистики является изучение закономерностей, которым подчиняются массовые случайные явления и процессы, с применением методов теории вероятностей.

Первая задача прикладной статистики — указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или специально поставленных экспериментов. Вторая задача прикладной статистики — разработать методы анализа статистических данных в зависимости от целей исследования. Сюда относятся оценка неизвестной вероятности события, оценка неизвестной функции распределения, оценка параметров распределения, оценка зависимости случайной величины от одной или нескольких случайных величин и т. д.

Итак, задача прикладной статистики заключается в разработке методов сбора и обработки статистических данных для получения научных и практических выводов. Основным методом изучения массовых случайных явлений в прикладной статистике является выборочный метод, суть которого состоит в том, что суждение обо всем множестве изучаемых объектов выносится на основе наблюдения за некоторой (возможно, относительно малой) частью. Неформальное описание выборочного метода дано в следующем параграфе, а необходимая формализация будет проведена позже, после введения понятия случайной величины.

Фундаментом прикладной статистики является математическая статистика. Прикладную статистику нельзя целиком относить к математике, поскольку она включает в себя две внематематические об-

ласти: методологию организации статистического исследования и организацию компьютерной обработки данных, в том числе разработку и использование баз данных, электронных таблиц, статистических программных продуктов и систем анализа данных.

1.2. Основные понятия выборочного метода: генеральная совокупность и выборка

Пусть требуется изучить, как в совокупности однородных объектов распределен некоторый качественный или количественный признак, характеризующий эти объекты. Например, имеется множество банковских заемщиков, качественным признаком каждого из них может служить пол, а количественным — годовой доход в рублях. Иногда проводят сплошное обследование, т. е. для каждого из объектов совокупности изучается интересующий признак. На практике, однако, сплошное обследование применяют сравнительно редко. Так, если население города очень большое, то провести сплошное обследование всех жителей физически невозможно. Или, например, если обследование прибора связано с его разрушением, требует больших материальных затрат, то в этом случае проводить сплошное обследование практически не имеет смысла. В таких ситуациях случайно отбирают из всей совокупности ограниченное число объектов и подвергают их изучению.

Определение 1. *Выборочной совокупностью, или просто выборкой, называют совокупность случайно отобранных объектов.*
При этом выборку осуществляют из генеральной совокупности.

Определение 2. *Генеральной совокупностью называют совокупность всех объектов, относительно которых предполагается делать выводы при изучении конкретной задачи.*

Генеральная совокупность состоит из всех объектов, которые имеют качества, свойства, интересующие исследователя.

Вопрос отбора объектов из генеральной совокупности отнюдь не является тривиальным, и от способа организации этого отбора зависит качество выборки. Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем признаке генеральной совокупности, необходимо, чтобы объекты выборки правильно его представляли. Другими словами, выборка должна правильно представлять пропорции генеральной совокупности. Это требование коротко формулируют так: выборка должна быть репрезентативной (представительной).

В силу закона больших чисел [1] можно утверждать, что выборка будет репрезентативной, если ее осуществить случайно: каждый объект выборки отобран из генеральной совокупности случайно, т. е. никаким объектам при отборе не отдается предпочтение.

Одним из наиболее известных исторических примеров нерепрезентативной выборки считается случай, происшедший во время президентских выборов в США в 1936 г. Журнал «Литерари Дайджест», успешно прогнозировавший события нескольких предшествующих выборов, ошибся в своих предсказаниях, разослав 10 млн пробных бюллетеней своим подписчикам, а также людям, выбранным по телефонным книгам всей страны, и людям из регистрационных списков автомобилей. В 25 % вернувшихся бюллетеней (почти 2,5 млн) голоса были распределены следующим образом:

1) 57 % отдавали предпочтение кандидату-республиканцу А. Лэндону;
2) 40 % выбрали действующего в то время президента-демократа Ф. Рузвельта.

На выборах же, как известно, победил Рузвельт, набрав более 60 % голосов. Ошибка «Литерари Дайджест» заключалась в следующем: желая увеличить репрезентативность выборки, работники журнала, которым было известно, что большинство их подписчиков считают себя республиканцами, расширили выборку за счет людей, выбранных из телефонных книг и регистрационных списков. Однако они не учли современных реалий и набрали еще больше республи-

канцев: во время Великой депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и высшего класса (т. е. большинство республиканцев, а не демократов).

Одна и та же выборка может рассматриваться как репрезентативная и как нерепрезентативная в зависимости от того, на какую генеральную совокупность исследователь желает распространить свои выводы.

П р и м е р. Выборка составлена по результатам периодического медицинского осмотра работников Богословского алюминиевого завода (выявление бронхолегочной патологии). Но если ставится задача исследования структуры бронхолегочной патологии жителей Свердловской области, то такую выборку следует считать нерепрезентативной. Однако при исследовании структуры бронхолегочной патологии работников алюминиевого производства в Российской Федерации та же самая выборка может считаться репрезентативной.

1.3. Понятие случайной величины

Строгое определение случайной величины требует привлечения понятийного аппарата теории функций вещественной переменной, но в рамках настоящего учебного пособия этого делать не нужно. Для изложения дальнейшего материала нам достаточно лишь общего понимания того, что собой представляет случайная величина, а потому мы ограничимся неформальным определением.

Определение 3. *Случайная величина — это величина, которая в результате испытания принимает одно и только одно возможное значение, наперед неизвестное и зависящее от случайных причин, которые заранее не могут быть учтены.*

П р и м е р 1. Число мальчиков из 100 новорожденных есть величина случайная, возможные значения которой: 0, 1, 2, . . . , 100.

П р и м е р 2. Дневная выручка магазина, выраженная в рублях.

П р и м е р 3. Среднесуточная температура в январе в Москве.

Будем далее обозначать случайные величины прописными буквами X, Y, Z, а их возможные значения — соответствующими строчными буквами x, y, z. Например, если случайная величина X имеет три возможных значения, то они будут обозначены так: x_1, x_2, x_3.

Разберем примеры 1 – 3. В первом из них случайная величина X могла принять одно из следующих возможных значений: 0, 1, 2, ..., 100. Эти значения отделены одно от другого промежутками, в которых нет возможных значений X. Таким образом, в этом примере случайная величина принимает отдельные, изолированные возможные значения. Во втором примере случайная величина также могла принимать только целочисленные неотрицательные значения, хотя ее границы точно неизвестны. В третьем примере случайная величина могла принять любое из значений промежутка (a, b). Здесь нельзя отделить одно возможное значение от другого промежутком, не содержащим возможных значений случайной величины.

Из сказанного можно сделать вывод о целесообразности различать случайные величины, принимающие лишь отдельные, изолированные значения, и случайные величины, возможные значения которых сплошь заполняют некоторый промежуток.

Определение 4. *Дискретной называют случайную величину, которая принимает отдельные, изолированные возможные значения с определенными вероятностями.*

Число возможных значений дискретной случайной величины может быть конечным или бесконечным.

Определение 5. *Непрерывной называют случайную величину, которая может принимать все значения из некоторого конечного или бесконечного промежутка.*

Очевидно, что число возможных значений непрерывной случайной величины бесконечно.

Для задания (определения) дискретной случайной величины (ДСВ) необходимо указать все принимаемые ею значения и соответствующие вероятности, т. е. ее закон распределения.

Определение 6. *Закон распределения дискретной случайной величины — соответствие между возможными значениями и их вероятностями.*

Обычно закон распределения ДСВ представляют в виде таблицы, первая строка которой содержит возможные значения, а вторая — их вероятности. Удобным способом представления закона распределения ДСВ является графический. При этом на оси абсцисс откладывают варианты x_i, а на оси ординат — соответствующие им вероятности p_i.

Задание закона распределения в виде таблицы требует перечисления всех значений случайной величины. Очевидно, что такой способ задания неприменим для непрерывных случайных величин, соответственно, необходимо дать общий способ задания любых типов случайных величин. С этой целью вводят функции распределения вероятностей случайной величины.

Пусть x — действительное число. Вероятность события, состоящего в том, что случайная величина X примет значение, меньшее x (т. е. вероятность события $X < x$), обозначим через $F_X(x)$. Разумеется, если x изменяется, то, вообще говоря, изменяется и $F_X(x)$, т. е. $F_X(x)$ есть функция от x.

Определение 7. *Функцией распределений случайной величины X называется функция $F_X(x)$, определяющая вероятность того, что случайная величина X в результате испытания примет значение, меньшее x, т. е. $P(X < x) = F_X(x)$.*

Геометрически это равенство можно истолковать так: $F_X(x)$ есть вероятность того, что случайная величина X примет значение, которое лежит на числовой оси левее точки x.

В терминах функции распределения можно дать более точное определение непрерывной случайной величины: случайную величину называют непрерывной, если ее функция распределения есть непрерывная, кусочно-дифференцируемая функция с непрерывной производной.

Другим способом определения непрерывной случайной величины является задание плотности распределения:

Определение 8. *Плотностью распределения вероятностей непрерывной случайной величины X называют функцию $f(x)$ — первую производную от функции распределения $F(x)$, т. е. $f(x) = F'(x)$.*

Зная плотность распределения, можно вычислить вероятность того, что непрерывная случайная величина примет значение, принадлежащее заданному интервалу. Правило вычисления дает следующее утверждение.

Утверждение 1. *Вероятность того, что непрерывная случайная величина X примет значение, принадлежащее интервалу (a, b), равна определенному интегралу от плотности распределения, взятому в пределах от a до b :*

$$P(a < X < b) = \int_a^b f(x)dx.$$

Напомним, что геометрический смысл определенного интеграла — площадь под кривой $y = f(x)$ в промежутке от a до b. Это утверждение позволяет раскрыть вероятностный смысл плотности распределения. Вероятность того, что случайная величина примет значение, принадлежащее интервалу $(x, x + \Delta)$, приближенно равна произведению плотности вероятности в точке x на длину интервала.

Зная плотность распределения, можно найти функцию распределения:

$$F_X(x) = \int_{-\infty}^{x} f(t)dt.$$

Подробно о свойствах функции и плотности распределения можно прочитать в [1, гл. 10–11].

1.4. Формализация понятий выборочного метода

Пусть проводятся наблюдения за случайной величиной X, распределение которой нам частично или полностью неизвестно. В математической статистике принято следующее определение:

Определение 9. *Генеральной совокупностью случайной величины X (или просто генеральной совокупностью X) называется множество возможных значений случайной величины X. Законом распределения (распределением) генеральной совокупности X называется закон распределения случайной величины X.*

Исходным материалом для изучения свойств генеральной совокупности (т. е. некоторой случайной величины) являются экспериментальные (статистические) данные, под которыми понимают значения случайной величины, полученные в результате повторений случайного эксперимента (наблюдений за случайной величиной).

Предполагается, что эксперимент хотя бы теоретически может быть повторен сколько угодно раз в одних и тех же условиях. Под словами «в одних и тех же условиях» будем понимать, что распределение случайной величины X_i, $i = 1, 2, \ldots$, заданной на множестве исходов i-го эксперимента, не зависит от номера испытания и совпадает с распределением генеральной совокупности X. В этом случае принято говорить о независимых повторных экспериментах (испытаниях) или о независимых повторных наблюдениях над случайной величиной.

Определение 10. *Случайной выборкой $\overrightarrow{X_n} = (X_1, \ldots, X_n)$ объема n из генеральной совокупности X называется набор из n независи-*

мых случайных величин $X_1, \ldots, X_n$, каждая из которых имеет то же распределение, что и случайная величина X.

При этом число n называют объемом случайной выборки, а случайные величины X_i — элементами случайной выборки.

Очевидно, что случайная выборка — объект абстрактный, в эксперименте не наблюдаемый. И вообще, случайной выборки в распоряжении исследователя быть не может, так как знание закона распределения случайной выборки равносильно знанию закона распределения генеральной совокупности X.

Выше мы говорили о данных, полученных в результате повторных наблюдений за случайной величиной. Оказывается, что для строгого обоснования статистических методов удобнее другой подход: вместо n-кратного наблюдения за случайной величиной X мы один раз наблюдаем за n случайными величинами $X_i,\ i = 1, 2, \ldots, n$, которые устроены так же, как исходная случайная величина X. Суть определения 10 в том, чтобы дать соответствующую формализацию этой идее.

В некотором смысле идея замены n-кратного наблюдения за одной случайной величиной X одним наблюдением за n случайных величин $X_i,\ i = 1, 2, \ldots, n$, близка к эргодической теории[1].

Что же реально есть у исследователя для получения конкретных конструктивных числовых оценок? У исследователя есть выборка (обратите внимание, что слово «случайная» мы здесь не пишем).

Определение 11. *Выборкой $\overrightarrow{x_n} = (x_1, \ldots, x_n)$ из генеральной совокупности X называется любое возможное значение случайной выборки $\overrightarrow{X_n}$.*

Число n характеризует объем выборки, а числа $x_1, ..., x_n$ представляют собой элементы выборки $\overrightarrow{x_n}$.

[1]Эргодическая гипотеза в статистической физике утверждает, что в состоянии равновесия для физических величин среднее значение по ансамблю равно среднему значению по времени.

Выборку $\overrightarrow{x_n}$ можно интерпретировать как совокупность n чисел $x_1, ..., x_n$, полученных в результате проведения n повторных независимых наблюдений над случайной величиной X.

Основой любых выводов о вероятностных свойствах генеральной совокупности X, т. е. статистических выводов, является выборочный метод, суть которого заключается в том, что свойства случайной величины X устанавливаются путем изучения тех же свойств на случайной выборке.

Задания для самостоятельной работы

1. В Москве проводится опрос населения с целью выяснить общественное мнение. Например, требуется узнать, где люди планируют провести летний отпуск. Приведите пример неправильной организации опроса, результатом которого будет нерепрезентативная выборка.

2. В чем отличие дискретной случайной величины от непрерывной? Приведите примеры той и другой случайной величины.

3. Бросают игральный кубик. Случайная величина X — число очков, выпавших на кубике. Составьте закон распределения случайной величины X.

4. Определено ли понятие плотности распределения для дискретных случайных величин? Если да, то как, если нет, то почему?

2. Сбор и первичная обработка данных

В главе разобраны отдельные кейсы, связанные с типичными ошибками, которые возникают на этапе сбора данных и внесения первичных данных в таблицы. Эти ошибки особенно часто встречаются, когда практикующие специалисты (врачи, психологи, экономисты) проводят исследование самостоятельно, не консультируясь со специалистами по статистическому анализу.

2.1. Шкалы измерений в статистике

Грамотное использование статистических методов обработки данных во многом зависит от четкого понимания исследователем того, как интерпретировать числа, внесенные в базу данных.

Пусть, например, проводится клиническое исследование и в базу данных внесены сведения о поле пациентов. Для удобства работы можно использовать кодировку: 0 — мужской пол, 1 — женский. Очевидно, что обозначение цифрами 0 и 1 соответственно лиц мужского и женского пола абсолютно произвольно, цифры можно было поменять местами, а можно для кодирования использовать другие цифры, например, 1 и 2. Мы, разумеется, не имеем в виду, что женщины стоят на ступеньку выше мужчин или мужчины значат больше, чем женщины. Таким образом, отдельным числам не соответствуют никакие эмпирические значения. В этом случае говорят о переменных, относящихся к *номинальной* шкале. В нашем примере рассматривается переменная с номинальной шкалой, имеющая две категории. Такая переменная имеет еще одно название — *дихотомическая*.

Такая же ситуация и с переменной «раса». Пусть в базе она имеет четыре значения: европеоидная раса — 1, негроидная — 2, восточноафриканская — 3, монголоидная — 4. Здесь также соответствие между числами и категориями расы не имеет никакого эмпирического значения, но, в отличие от пола, эта переменная не является дихотомической, у нее четыре категории вместо двух. Возможности

обработки переменных, относящихся к номинальной шкале, очень ограничены. Можно провести только частотный анализ таких переменных. К примеру, расчет среднего значения для расы совершенно бессмыслен. Переменные, относящиеся к номинальной шкале, часто используются для группировки, с помощью которой совокупная выборка разбивается по категориям этих переменных. В частичных выборках проводятся одинаковые статистические тесты, результаты которых затем сравниваются друг с другом.

В качестве следующего примера рассмотрим переменную «стадия хронической болезни почек» (ХБП). Современная классификация основана на двух показателях — скорости клубочковой фильтрации (СКФ) и признаках почечного повреждения (протеинурия, альбуминурия). В зависимости от их сочетания выделяют пять стадий хронической болезни почек, при этом третью стадию разделяют на две: А и Б (табл. 1).

Таблица 1

Классификация хронической болезни почек

Код	Стадия ХБП	Признаки	СКФ, мл/мин/1.73 м^2
1	1	Признаки нефропатии	> 90
2	2	Признаки нефропатии	60 — 89
3	3А	Умеренное снижение СКФ	45 — 59
4	3Б	Выраженное снижение СКФ	30 — 44
5	4	Тяжелое снижение СКФ	15 — 29
6	5	Терминальная почечная недостаточность	< 15

Здесь кодовым цифрам присваивается эмпирическое значение в том порядке, в котором они расположены в списке. Код для перемен-

ной ХБП отсортирован в порядке возрастания стадии ХБП: пациент с терминальной хронической почечной недостаточностью находится в более худшем состоянии, нежели пациент с тяжелым снижением СКФ, а пациент с тяжелым снижением СКФ — в более худшем состоянии, чем пациент с выраженным снижением СКФ, и т. д. Такие переменные, для которых используются численные значения, соответствующие постепенному изменению признака, относятся к *порядковой* шкале.

Однако содержательный смысл этих переменных не зависит от разницы между соседними численными значениями. Так, несмотря на то что разница между значениями кодовых чисел для стадии 5 и стадии 4, а также для стадии 4 и стадии 3Б в обоих случаях равна единице, нельзя утверждать, что фактическое различие между состояниями пациентов стадии 5 и стадии 4, а также пациентов стадии 4 и стадии 3Б одинаково.

Классическим примером, где переменная относится к порядковой шкале, являются всевозможные рейтинги, а также места, занятые участниками соревнований.

Кроме частотного анализа, переменные с порядковой шкалой допускают также вычисление определенных статистических характеристик, таких как медианы. Если должна быть исследована связь с другими переменными такого рода, то для этой цели можно использовать коэффициент ранговой корреляции Спирмена или Кендалла. Для сравнения различных выборок переменных, относящихся к порядковой шкале, могут применяться непараметрические тесты, формулы которых оперируют рангами.

Рассмотрим теперь коэффициент интеллекта IQ. Его абсолютные значения отображают порядковое отношение между респондентами, и разница между двумя значениями также имеет содержательный смысл. Например, если у Владимира IQ равен 90, у Ивана — 120, а у Владислава — 150, можно сказать, что Иван настолько же интеллектуальнее Владимира, насколько Владислав интеллектуальнее

Ивана (а именно на 30 единиц). Однако тот факт, что у Владислава значение IQ в 1.25 раза больше, чем у Ивана, не позволяет на основании определения IQ сделать вывод, что Владислав на 25 % умнее Ивана.

Такие переменные, у которых разность (интервал) между двумя величинами имеет содержательный смысл, но отношение величин этого смысла лишено, относятся к *интервальной* шкале. Они могут обрабатываться любыми статистическим методами без ограничений. Так, к примеру, среднее значение является полноценным статистическим показателем для характеристики таких переменных.

Классическим примером, где переменная относится к интервальной шкале, являются шкалы температур Цельсия и Фаренгейта. Ноль, формально присутствующий в этих шкалах, выбран весьма произвольно.

Наконец, мы подошли к статистической шкале *отношений*, где отношение двух величин приобретает содержательный смысл. Примером переменной, относящейся к такой шкале, является возраст: так, если Владиславу 25 лет, а Ивану 50, можно сказать, что Владислав вдвое младше Ивана. Шкала, к которой относятся данные, называется шкалой отношений. Эта шкала включает все интервальные переменные, которые имеют абсолютную нулевую точку. Поэтому переменные, относящиеся к интервальной шкале, как правило, имеют и шкалу отношений. С помощью таких шкал могут быть измерены масса, длина, концентрация. Шкала Кельвина (температуры, отсчитанные от абсолютного нуля, с выбранной по соглашению специалистов единицей измерения Кельвин) является примером шкалы отношений.

В итоге интерпретация данных, внесенных в базу, должна проводиться в соответствии с тем, к какой шкале данные были отнесены. Выделяются четыре вида шкал:

1. Номинальная шкала. Числа, хранимые в базе данных, являются условным кодом, например, чистота кредитной истории (1 — есть

случаи невозврата кредитов, 0 — нет таких случаев).

2. Порядковая шкала. Числа, хранимые в базе данных, выражают степень развития признака, например, уровень компетенций сотрудника (1 — junior, 2 — middle, 3 — senior).

3. Интервальная шкала. Числа, хранимые в базе данных, характеризуют физическую и/или экономическую величину в единицах ее измерения, при этом можно оценивать, на сколько одно значение больше другого, но нельзя оценивать во сколько раз одно значение больше другого (например, температура тела в градусах Цельсия, уровень IQ в баллах).

4. Шкала отношений. Числа, хранимые в базе данных, характеризуют физическую и/или экономическую величину в единицах ее измерения, при этом можно оценивать, во сколько раз одно значение больше другого (например, стоимость в рублях, длина в метрах, вес в килограммах).

Ранее мы выяснили, что переменные, относящиеся к номинальной шкале, допускают весьма ограниченные возможности для проведения анализа. Исключение в некоторых ситуациях составляют дихотомические переменные. Для них можно, по крайней мере, определять ранговую корреляцию. Если, например, обнаруживается корреляция коэффициента интеллекта с полом, то положительный коэффициент корреляции означает, что женщины интеллектуальнее, чем мужчины. Однако если переменные, относящиеся к номинальной шкале, не являются дихотомическими, вычисление коэффициентов ранговой корреляции не имеет смысла. Так, например, если одна из переменных выражает код профессии (1 — машинист крана, 2 — слесарь и т. д.), а другая — степень артериальной гипертензии, то вычисление коэффициента корреляции лишено смысла.

Правильное применение статистических методов обработки данных во многом зависит от четкого понимания исследователем того, в какой статистической шкале они представлены. Непонимание этого может привести к тому, что исследователь получит результа-

ты, которые не отражают действительное положение вещей, и сделает неправильные выводы. Именно поэтому понимание того, в какой шкале представлены статистические данные, является одним из необходимых условий успешного статистического анализа.

2.2. Работа с пропущенными значениями

На практике в реальных данных очень часто встречаются пропуски (англ. *missing data*). Например, при проведении клинических исследований некоторым пациентам не назначают анализы. При проведении социологических опросов респонденты могут отказаться отвечать на некоторые вопросы. Причинами пропусков могут быть ошибки ввода данных, утеря или сокрытие информации.

Начнем с примера неправильной обработки пропущенных значений, а далее разберем подходы к работе с пропущенными значениями.

Рассмотрим факторы, влияющие на риск развития острого послеоперационного делирия после эндопротезирования тазобедренного сустава. Гемостаз после этой операции наступает обычно на вторые сутки, а потому важным лабораторным показателем, позволяющим оценить состояние пациента в послеоперационный период, является уровень гемоглобина на вторые сутки. Если объем кровопотери (видимой интраоперационной и дренажной в первые двое суток) низкий и пациент чувствует себя хорошо, данный анализ иногда не проводят. Соответственно, ничего неизвестно про уровень гемоглобина у этих пациентов.

Исследователь рассуждает: «Ничего — это, как мы знаем со школы, ноль» — и ставит нули во всех ячейках, где не было данных о гемоглобине на вторые сутки. При проведении расчетов алгоритмы, встроенные в пакеты программ, воспринимают эти нули как обычные данные, и все оценки получаются смещенными, а статистические выводы — неверными. В частности, на основе таких неверных данных можно прийти к выводу, что риски развития послеопераци-

онных осложнений не связаны с уровенем гемоглобина на вторые сутки.

Можно ли исправить такую таблицу с данными? Да. Так как очевидно, что нулевых значений в столбце «Уровень гемоглобина» быть не должно, можно автоматически (пакеты программ умеют делать такие преобразования) все ячейки с нулями сделать пустыми. Далее можно исключить таких пациентов и использовать для проведения анализа только верные данные (традиционный подход) или применить специальные методы анализа пропущенных значений (современный подход).

П р и м е р. Проводится мониторирование интраоперационного систолического давления. Во время операции у пациента может возникнуть асистолия — прекращение деятельности сердца с исчезновением биоэлектрической активности. Однако в результате реанимационных действий сердце может быть снова запущено. У части пациентов нет данных о систолическом давлении[2], а потому про этот показатель ничего неизвестно. Снова исследователь рассуждает: «Ничего — это ноль» — и ставит нули во всех ячейках, где не было данных.

Исправить ошибки в этом примере гораздо сложнее, чем в вышеописанном, так как здесь смешиваются данные о реальных случаях асистолии и нули, внесенные по ошибке. Здесь особенно важно различать пропущенные значения и нулевые.

Как же обрабатывать пропущенные значения? Методы обработки зависят от типа пропусков: полностью случайные, случайные и неслучайные пропуски [2].

1. MCAR (Missing Completely at Random). Полностью случайные пропуски имеют место в тех случаях, когда подвыборка значений по переменной(-ым), подлежащей изучению, по-прежнему является моделью генеральной совокупности. Приведем пример данных, содер-

[2]Например, решение заносить интраоперационное систолическое давление в протокол было принято в середине исследования или в многоцентровом исследовании лишь в двух клиниках фиксировали этот показатель.

жащих пропуски MCAR-типа. Проводится опрос населения с целью выяснения политических предпочтений. У некоторых респондентов данные о политических предпочтениях пропущены, однако эти данные не зависят от других переменных (например, образование, возраст, уровень доходов и т. д.). Кроме того, вероятность пропуска не зависит от значения самой переменной, т. е. не возникает ситуаций, когда респонденты с определенной политической позицией чаще других не дают ответа на соответствующий вопрос. Причина пропусков случайна: респондент не понял суть вопроса, невнимательно читал анкету и не заметил вопрос.

Выбор модели полностью случайных пропусков — это единственное допущение, которое можно проверить эмпирически. Что касается случайных и неслучайных пропусков, то соответствующие допущения невозможно проверить на основании имеющегося массива.

2. MAR (Missing at Random). Случайные пропуски имеют место в тех случаях, когда вероятность пропуска зависит от значений других переменных, но не зависит от самих пропущенных значений. Приведем пример пропусков MAR-типа. Пожилые респонденты более склонны скрывать свои политические предпочтения, чем молодые люди, но внутри старшей возрастной группы пропуски распределены случайно. Иными словами, скрытность пожилых респондентов не зависит от их политической ориентации (консерваторы, либералы и т. д.), в то же время молодые респонденты открыто заявляют о любых своих политических предпочтениях.

В этом случае возможно смещение результатов оценивания параметров. Например, если молодые респонденты более либерально настроены, а пожилые — более консервативно (но скрывают это), то оценка среднего значения будет смещаться в сторону либеральных настроений. Если же политические предпочтения от возраста не зависят, то смещения результатов оценивания не произойдет.

3. MNAR (Missing Not at Random). Неслучайные пропуски имеют место в тех случаях, когда вероятность пропуска зависит от самих

пропущенных значений. Например, люди с левыми политическими взглядами более склонны скрывать свои политические предпочтения. Такие пропуски обязательно вносят систематические ошибки в результаты анализа.

Методы обработки пропущенных значений делят на традиционные и современные. К первым относят построчное и попарное удаление наблюдений, замещение средним и замещение с использованием регрессии. Ко вторым — замещение посредством оценки максимального правдоподобия, множественное замещение, селективную модель и модель смешанных паттернов.

При построчном удалении из массива исключаются любые наблюдения, в значениях переменных которых присутствует хотя бы один пропуск. Такой подход может использоваться только при MCAR. Разумеется, его использование ведет к снижению статистической мощности.

Попарное удаление заключается в том, что в каждом конкретном случае анализа из него удаляются только те наблюдения, в которых присутствует хотя бы один пропуск по релевантным для данного анализа переменным. Как и при построчном удалении, попарное удаление подходит исключительно в случае MCAR. При этом создается дополнительная проблема, состоящая в том, что для проверки различных гипотез одного исследовательского проекта используются отличающиеся подвыбороки.

Использование замещения пропущенных значений выборочным средним не рекомендуется в любом случае. Это связано с тем, что такой способ обработки пропусков ведет к смещенным оценкам (поскольку уменьшает дисперсию переменных и ковариацию между ними), независимо от используемого допущения. Исключением является оценка самого среднего значения.

Замещение с использованием регрессии (вместо пропущенных используются предсказанные значения) имеет схожие с предыдущим способом проблемы. Так, получаемые значения попадают в один пат-

терн, отвечающий регрессионной прямой, что вносит смещение в оценки дисперсии и ковариации. Частично эта проблема может быть решена посредством добавления случайных остаточных значений к предсказанным величинам. В целом же данный способ замещения подходит при MAR [3, 4].

Замещение посредством оценки максимального правдоподобия представляет собой итерационный процесс, каждая итерация которого включает два этапа: 1) средние значения переменных и ковариационная матрица используются для расчета регрессионных уравнений, предсказывающих пропущенные значения; 2) полученный массив используется для расчета обновленных средних значений и ковариационной матрицы. Эти этапы повторяются до тех пор, пока средние значения и ковариационная матрица не перестанут изменяться. Можно сказать, что в данном случае используется последовательный подход.

В случае множественного замещения создается набор альтернативных массивов данных (от пяти и более), в которых пропущенные значения замещаются предсказанными посредством регрессии с добавлением случайного элемента. После этого средние значения переменных и ковариационные матрицы обобщаются с целью получения итоговой оценки. Этот подход можно назвать конкурентным. Как замещение посредством максимального правдоподобия, так и множественное замещение подходят для ситуаций, в которых исследователь исходит из MCAR- или MAR-допущений. Для случаев, когда делается НП-допущение, предназначены селективная модель и модель смешанных паттернов.

Селективная модель соединяет основное регрессионное уравнение с дополнительным регрессионным уравнением, предсказывающим вероятность ответов. Две части модели связываются посредством скоррелированных остатков (англ. *correlated residuals*), и эта связь является механизмом, с помощью которого корректируются смещения модели, связанные с пропусками.

Модель смешанных паттернов предполагает формирование подгрупп наблюдений, в которых обнаруживаются одинаковые паттерны пропусков данных, с дальнейшей оценкой интересующих параметров в каждой подгруппе. После чего находится средневзвешенная оценка этих параметров.

В отличие от современных методов обработки пропущенных значений при MCAR- или MAR-допущениях, методы обработки при MNAR-допущении не получили достаточно широкого признания [5].

2.3. Возможность автоматического вычисления значений в базе

Составляя таблицы с данными, необходимо заранее спланировать, какие данные будут первичными, а какие вычисляемыми.

П р и м е р. Исследуется выживаемость пациентов после аллотрансплантации почки. После того как было прооперировано 100 пациентов, исследователь начал вносить сведения в базу данных из архива, при этом завел столбцы «Время жизни» и «Статус пациента». Статус пациента — бинарная величина, равная 0, если пациент жив, и равная 1, если пациент не жив. Время жизни — количество месяцев, которые пациент прожил после операции либо до своей смерти (в этом случае в столбце «Статус пациента» стоит 1), либо до текущего дня (в этом случае в столбце «Статус пациента» стоит 0). Пациенты периодически приходят на профилактический осмотр, и каждый раз врач вручную пересчитывает продолжительность жизни пациента после трансплантации. Если пациент, находящийся на учете, умирает, данные о его смерти поступают врачу, и он вносит эти сведения в базу. Такой подход очень затратный в плане поддержания базы в актуальном состоянии и требует ручных расчетов, а потому подвержен ошибкам.

Верным подходом в данном случае является занесение в базу данных даты операции и даты последнего визита. Время жизни пациента после трансплантации вычисляется при этом автоматически.

2.4. Суррогатные группировки

Важно заносить в базу первичные, натурные, данные. При неверном подходе данные группируются в некоторые классы, а потом в базу заносится лишь факт принадлежности к соответствующему классу. Мы называем такие преобразования исходных данных суррогатными группировками. При суррогатных группировках теряется информация.

П р и м е р. У пациентов измеряется уровень гемоглобина в граммах на литр. Исследователь разбивает пациентов на две группы: гемоглобин больше или равен 90 г/л — норма, менее 90 г/л — анемия, и вносит в базу принадлежность пациента к группе 0 и 1 соответственно. Далее исследователь ставит задачу сравнения этих двух групп по распространенности тех или иных осложнений (ишемических, септических и т. д.).

При группировке теряется существенная часть информации. Действительно, попасть в группу 0 могут пациент А. с уровнем гемоглобина 90 г/л и Б. с уровнем гемоглобина 120 г/л, а в группу 1 — пациенты В. и Г. с уровнем гемоглобина 87 г/л и 70 г/л соответственно.

Очевидно, что состояние пациентов А. и Б. весьма отличается (по уровню гемоглобина), но программа, проводящая статистический анализ, их воспринимает как одинаковых. С пациентами В. и Г. аналогично. Несмотря на то что пациенты А. и В., по сути, весьма похожи, в базе они относятся к различным группам.

Разница натурных показателей у А. и Б. равна: $120 - 90 = 30$ г/л, а разница натурных показателей у Б. и В. равна: $90 - 87 = 3$ г/л. Таким образом, принадлежность к этим суррогатным группам не отражает реальность, а потому статистические связи, которые имеются между уровнем гемоглобина и риском осложнений, могут разрушиться (или ослабеть), и статистические выводы будут неверными.

При суррогатных группировках происходит потеря информации. Сложным вопросом остается и выбор точки разбиения. В данном

примере выбор границы 90 г/л, вообще говоря, не был чем-либо обоснован.

2.5. Ошибки в формировании групп

Проводя анализ данных, необходимо следить за тем, чтобы все величины имели содержательный смысл в контексте решаемой задачи, группы формировались на основе действительно общего признака. Приведем четыре содержательных примера.

П р и м е р 1. Изучается распространенность хронического простатита в группе пациентов старше 40 лет. Исследователь берет базу всех пациентов урологического отделения, делит количество пациентов с установленным диагнозом «хронический простатит» на количество всех пациентов старше 40. Полученная таким образом оценка распространенности оказывается почти в два раза ниже, чем описано в литературных источниках. Этот результат связывают с успешно проводимой оптимизацией здравоохранения в отдельно взятом регионе и реализацией программы «Мужское здоровье». При этом исследователь не учитывает, что половина пациентов в базе — женщины.

П р и м е р 2. Изучается функция пересаженной почки. Исследователь оценивает функцию почки по уровню креатинина и скорости клубочковой фильтрации, находит средние и 95 % доверительный интервалы для этих величин. Однако он не учитывает, что часть пациентов после трансплантации вернулись на гемодиализ. Оценка уровня креатинина у таких пациентов — это лишь оценка качества диализа, но никак не функции погибшей почки.

В этих примерах неверно сформированы группы. Распространенность простатита нужно искать только среди мужчин, а функцию почки оценивать (находя средние и 95 % доверительный интервалы) только в группе пациентов с функционирующей пересаженной почкой.

П р и м е р 3. Изучается влияние вредного стажа в горячем цеху металлургического комбината на распространенность артериальной

гипертензии (АГ). Обследовано 100 рабочих. Исследователь формирует стажевые группы (табл. 2).

Таблица 2

Влияние стажа работы в горячем цеху на распространенность артериальной гипертензии

Группа	Стаж, лет	Распространенность АГ, %	Численность чел.
А	0 — 9	15	60
Б	10 — 19	40	30
В	20 и более	50	10

Далее исследование разделено на этапы:

- На первом этапе рабочие группы А сравниваются с остальными 40 рабочими, т. е. рассматривается две (не три!) группы: со стажем до 9 лет (группа А) и от 10 лет (остальные). В группе А распространенность АГ оказывается значимо ниже ($p = 0.003$, Хи-квадрат).
- На втором этапе также рабочие группы Б сравниваются с остальными 70 рабочими. В группе Б распространенность АГ оказывается значимо выше, чем у остальных рабочих ($p = 0.037$, Хи-квадрат).
- На третьем этапе также сравниваются две группы: группа В и остальные 90 рабочих. В силу малочисленности этой группы (кто-то не доживает, кого-то выводят из производства) показатели АГ здесь численно выше, чем у остальных рабочих, но незначимо отличаются от показателей распространенности АГ у остальных рабочих ($p = 0.069$, Хи-квадрат).

На основании этих расчетов делается вывод, что наиболее опасен для здоровья стаж работы длительностью от 10 до 20 лет, а

потом организм адаптируется и риски уменьшаются. Вывод странный, поскольку адаптационные способности организма снижаются с возрастом. Разберем, почему так получилось.

На первом этапе все было верно. Казалось бы, задача второго этапа абсолютно аналогична задаче первого этапа (формулировка почти дословно совпадает), но это не так. «Остальные рабочие», с которыми производится сравнение, *не образуют однородную группу сравнения*! Сюда входят те, у кого стаж менее 10 лет, и те, у кого стаж от 20 и выше. Выделение подобной группы лишено смысла, результаты формально проведенного сравнения не могут быть проинтерпретированы с клинической точки зрения.

Действительно, результат, полученный на втором этапе, звучит так: «Вредный стаж длительностью от 10 до 20 лет более опасен, чем стаж (аналогично вредности) до 10 лет или *более 20*». Отсюда, в частности, следует, что работать в горячем цеху 30 лет не столь вредно, как 15.

То что распространенность АГ в группе Б выше, чем среди *остальных*, — это математический артефакт, причина которого в следующем. В эту группу из *остальных* 70 чел. по большей части входят рабочие из группы А (60 чел.), где распространенность АГ составляет всего 15 %, и небольшое количество рабочих из группы В, где распространенность АГ составляет 50 %. Взвешенное среднее равно 20: 0.15 + 10 0.5 = 0,2, т. е. 20 %.

На третьем этапе не было найдено статистически значимое отличие — это тоже математический артефакт, причина которого в малой численности группы: рабочих со стажем более 20 лет было всего 10 человек.

Если статистические выводы противоречат интуиции и опыту клинициста, необходимо проверить, верными ли были дизайн исследования и выбранные методы анализа. Для решения данной задачи следовало бы использовать логистическую регрессию, а не разбиение на отдельные группы.

Задания для самостоятельной работы

1. Известно, что у представителей негроидной расы нормальный уровень креатинина сыворотки крови выше, чем у других рас. Это, в частности, учитывается при определении скорости клубочковой фильтрации. Исследователь, желая выяснить связь расы, закодированной, как описано в п. 2.1, с уровнем креатинина, нашел коэффициент корреляции Пирсона. Коэффициент оказался статистически незначимо отличным от нуля. На основании этого исследователь сделал вывод, что уровень креатинина не связан с расами. Нет ли в этом анализе ошибки? Если да, то в чем она состоит? Какой альтернативный подход можно предложить?

2. Верны ли следующие утверждения о пропущенных данных?

А. Замена пропущенных данных (англ. *imputation*) на сгенерированные данные — это на самом деле просто выдумывание данных для искусственного повышения значимости результатов. Лучше не включать в анализ строки с пропущенными данными, чем проводить искусственные вставки.

Б. Недостающие данные можно заменить на среднее или медианное значение. Это не сделает оценки смещенными, но зато уменьшит стандартную ошибку среднего и повысит мощность используемых критериев.

В. Отсутствие данных на самом деле не проблема, если требуется провести простой тест, такой как Хи-квадрат и t-тест.

Г. Худшее, к чему приводят пропущенные данные, — это уменьшение размера выборки и уменьшение мощности.

3. У исследователя есть данные о росте и весе пациентов. Желая изучить связь между уровнем глюкозы натощак и ожирением пациентов, исследователь определяет группы согласно значениям индекса массы тела (ИМТ) (см. таблицу).

Группа	ИМТ	Описание
1	18.5 — 25	Нормальный вес
2	25 — 30	Избыточный вес
3	30 — 35	Ожирение I степени
4	35 — 40	Ожирение II степени
5	Более 40	Ожирение III степени

Не происходит ли потери информации при переходе от объективной величины ИМТ[3] к группам ожирения? Не является ли это примером суррогатной группировки?

[3]Индекс массы тела — величина, позволяющая оценить степень соответствия массы человека и его роста и тем самым косвенно судить о том, является ли масса недостаточной, нормальной или избыточной. Важен при определении показаний для необходимости лечения. Индекс массы тела рассчитывается по формуле: $I = m/h^2$, где m — масса тела в килограммах, h — рост в метрах, и измеряется в кг/м^2.

3. Проверка статистических гипотез

3.1. Понятия статистической гипотезы, ошибок первого и второго рода

Пусть изучается генеральная совокупность X, пусть $\overrightarrow{X_n} = (X_1, ..., X_n)$ — случайная выборка из генеральной совокупности X, а $\overrightarrow{x_n} = (x_1, ..., x_n)$ — выборка. Имея выборку, можно выдвинуть несколько взаимоисключающих гипотез о распределении генеральной совокупности, одну из которых следует предпочесть остальным.

Определение 12. *Предположение о распределении генеральной совокупности X или параметрах этого распределения называется статистической гипотезой.*

Приведем примеры статистических гипотез: генеральная совокупность распределена по нормальному закону; математическое ожидание генеральной совокупности равно 100; две случайные величины не коррелированы. Эти гипотезы вытекают из следующих содержательных задач: правда ли, что среднемесячный товарооборот с соседним государством составляет 500 млн руб.; правда ли, что повышение деловой активности в Европе приводит к росту цен на российскую нефть. Предположение, состоящее в том, что на Марсе есть жизнь, не является примером статистической гипотезы.

Наряду с основной гипотезой рассматривается противоречащая ей гипотеза. Нулевой (основной) называют выдвинутую гипотезу, ее обычно обозначают как H_0. Конкурирующей (альтернативной) называют гипотезу, которая противоречит нулевой, ее обычно обозначают как H_1.

Выдвинутая гипотеза нуждается в проверке на основе наблюдаемых значений (выборки). Под процедурой проверки статистических гипотез понимают последовательность действий, позволяющих с той или иной степенью достоверности подтвердить или опровергнуть утверждение гипотезы. Отметим, что отвергая основную гипотезу, мы отдаем предпочтение альтернативной (табл. 3).

Таблица 3

Проверка статистических гипотез

		Гипотеза, верная на самом деле	
		H_0	H_1
Принятая гипотеза	H_0	Верное решение	Ошибка II рода
	H_1	Ошибка I рода	Верное решение

Как видно из табл. 3, в результате проверки возможно принятие двух верных решений и двух ошибочных:

1) гипотеза H_0 верна и ее приняли в результате проверки;
2) гипотеза H_0 неверна и ее отвергли в результате проверки, приняв H_1;
3) гипотеза H_0 верна, но в ходе проверки ее ошибочно отвергли, приняв H_1;
4) гипотеза H_0 неверна, но в ходе проверки ее ошибочно приняли.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза H_0. Ошибка второго рода состоит в том, что будет принята неправильная гипотеза H_0.

В основе проверки статистических гипотез лежит принцип практической невозможности маловероятных событий, который гласит: «В единичном испытании маловероятное событие не должно появиться». Проиллюстрируем это примером.

П р и м е р. В ящике лежат два шара, цвет которых неизвестен. Исследователь выдвинул гипотезу, что шары разных цветов и, не глядя в ящик, достал шар — он оказался белым. Шар вернули обратно в ящик. Эксперимент повторили, шар снова оказался белым, — и так 100 раз. Вывод напрашивается сам — оба шара белые. Разумеется, есть вероятность того, что один шар был черный, но вероятность эта равна 2^{-100}, и поскольку она крайне мала, исследователь отверг гипотезу как противоречащую экспериментальным данным.

3.2. Основные этапы проверки статистических гипотез

Рассмотрим основные этапы проверки гипотез, иллюстрируя их примером.

П р и м е р. В открытом море перекладывают груз с одного корабля на другой. Колесный робот ездит по палубе первого корабля и манипулятором («механической рукой») захватывает контейнеры с палубы второго корабля, перекладывая их к себе. Первоначально робот был хорошо откалиброван, но в процессе эксплуатации мог испортиться.

Из официальной документации известно, что сила, с которой манипулятор хватает груз, распределена по закону, близкому к нормальному, и составляет в среднем 600 Н. Ясно, что если прилагаемая сила будет много меньше 600 Н, то контейнер выскользнет из манипулятора и утонет в море. Если прилагаемая сила будет много больше 600 Н, то контейнер погнется.

На ежегодном техобслуживании проведено 25 замеров силы манипулятора. Результаты представлены в таблице. Аналитики желают проверить, согласуются ли данные фактических замеров с заявленными в документации характеристиками. При этом отклонения и в ту и в другую сторону опасны.

624	598	609	592	588
578	598	604	616	628
634	605	590	628	632
584	627	612	606	620
641	585	641	637	654

Заявленное значение силы составляет: $m_0 = 600$. Среднее значение силы по результатам 25 измерений составляет: $m_1 = 613$, среднее квадратическое отклонение $\sigma = 20.97$, минимум составляет 578, максимум — 654.

Очевидно, что расчетное среднее больше 600, но возникает вопрос, насколько это значимо. Возможно, эта разница в 13 находится, как говорят, «в пределах статистической погрешности».

Этап 1. Формулировка основной гипотезы H_0 и альтернативной гипотезы H_1.

Гипотеза H_0 состоит в том, что $m_0 = m_1$. Независимый аналитик склонен считать, что имеющее место отклонение в развиваемой силе (m_1 вместо m_0) объясняется случайными факторами: качка в море, порывы ветра и т. п. Иными словами, он исходит из того, что H_0 верна.

В качестве альтернативной гипотезы аналитик выбирает $m_0 \neq m_1$. С содержательной точки зрения альтернативная гипотеза означает, что средняя сила, развиваемая манипулятором, отличается от 600 (возможно, в большую, а возможно, и в меньшую сторону[4]).

Этап 2. Выбор статистического критерия.

Ранее мы ввели в рассмотрение $\overrightarrow{X_n}$; еще раз обратим внимание на то, что $\overrightarrow{X_n}$ — это математическая абстракция, виртуальный объект, который удобно использовать для построения теории и формализации правил проверки гипотез. $\overrightarrow{X_n}$ — векторная случайная величина, в реальности нам доступна лишь одна ее реализация: $\overrightarrow{x_n}$ — числовой вектор.

Рассмотрим специально подобранную вещественнозначную функцию $\varphi(\overrightarrow{X_n})$. Будучи функцией от случайных величин, φ сама является случайной величиной, а значит, имеет функцию распределение. Выбирать вид φ надо так, чтобы ее точное или приближенное распределение было известно. Эта функция потребуется для проверки нулевой гипотезы.

Определение 13. *Статистическим критерием (или просто кри-*

[4]Последнее замечание окажется важно, когда мы будем строить критическую область. В данном примере она будет двусторонняя.

терием) называют случайную величину $\varphi(\overrightarrow{X_n})$, *которая служит для проверки нулевой гипотезы.*

Подбор вида функции φ абсолютно неформальная, нестандартная задача. Например, если проверяют гипотезу о равенстве дисперсий s_1 и s_2 в двух нормальных генеральных совокупностях, то в качестве φ выбирают отношение исправленных выборочных дисперсий: $\varphi = s_1/s_2$. Поскольку в различных опытах выборочные дисперсии принимают различные значения, наперед неизвестно какие, то и их отношение будет величиной случайной. Можно показать, что φ будет распределена по закону Фишера–Снедекора. При проверке гипотезы о равенстве нулю коэффициента корреляции полагают $\varphi = r\frac{\sqrt{n-2}}{1-r^2}$, которая при справедливости нулевой гипотезы имеет распределение Стьюдента с $n-2$ степенями свободы (здесь n — объем выборки, r — значение выборочного коэффициента корреляции). Эти примеры показывают, что общего правила здесь нет; но для большинства часто решаемых задач (сравнение средних, проверка на нормальность и т. д.) правила выбора функции φ известны.

Аналитик мог бы рассмотреть функцию такого вида:

$$\varphi = \frac{X_1 + X_2 + ... + X_n}{n} - 600,$$

которая распределена по нормальному закону и при условии справедливости нулевой гипотезы имеет математическое ожидание, равное нулю, здесь $n = 25$ — число измерений. Содержательный смысл этой величины — на сколько Ньютонов средняя по 25 измерениям сила оказалась больше (или, возможно, меньше) стандартного уровня в 600 Н. Величина эта случайная, потому что при иных обстоятельствах (ветер дул в другую сторону, шторм) или для других 25 измерений эта величина могла оказаться другой.

Исходя из математических соображений аналитик, однако, рас-

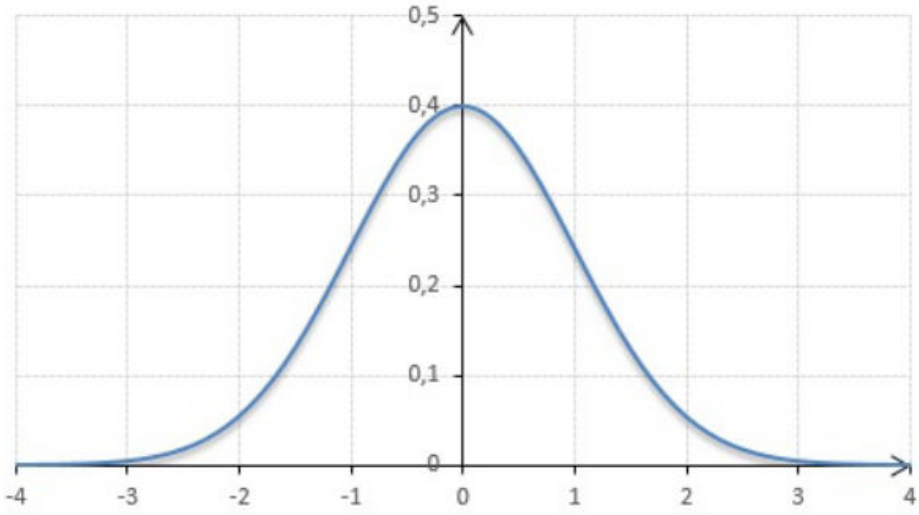


Рис. 1. Плотность нормального распределения

сматривает чуть более сложную величину:

$$\varphi = \frac{X_1 + X_2 + ... + X_n - 600n}{n\sigma}.$$

График плотности этой случайной величины показан на рис. 1. Как видно из рис. 1, значения в окрестности нуля случайная величина φ принимает чаще, а большие отклонения от нуля маловероятны.

Если бы в результате 25 измерений средняя сила составила 720 Н или 380 Н, иными словами, если бы отклонение средней силы от стандартного значения составило 120 Н или –120 Н, то аналитик сказал бы: «Если робот действительно был бы хорошо откалиброван, то такого сильного отклонения не наблюдалось бы, следовательно, гипотеза о том, что робот хорошо откалиброван, неверна». Или он мог сказать так: «Вероятность получить столь значительные отклонения крайне мала, а значит, на то была причина; и причина эта — износ деталей робота. Пора его ремонтировать».

Этап 3. Выбор уровня значимости. В предыдущем абзаце мы употребили оборот «вероятность крайне мала». Но что значит ма-

ла? Является ли вероятность 0.1 малой? И если нет, то является ли малой вероятность 0.001? Однозначного ответа здесь нет, более того, ответ на этот вопрос лежит вне компетенции математиков. В каждой предметной области практикующие специалисты сами решают, начиная с какого уровня вероятность считать малой, а соответствующие отклонения маловероятными.

С одной стороны, уменьшение порогового значения (например, до 10^{-6}) увеличивает опасность эксплуатации разболтанного робота, но зато при этом не требуются частые ремонты и без того нормально работающих роботов. С другой стороны, увеличение порогового значения (например, до 10^{-1}) снижает риск эксплуатации плохо функционирующего оборудования, но вместе с тем возникает необходимость в частых ненужных наладках нормально работающих роботов.

Исследователь должен определить для себя вероятность отправки на (ненужный) ремонт нормального робота, взамен он получит некоторую уверенность в том, что на корабль не поставят разболтанного робота. В ходе этих рассуждений мы пришли к понятию уровня значимости.

Определение 14. *Уровнем значимости называется вероятность ошибки первого рода, т. е. вероятность ошибочно отклонить верную нулевую гипотезу.*

Уровень значимости обычно обозначают буквой α и берут малым, например, $\alpha = 0.05$ или $\alpha = 0.01$.

Этап 4. Построение критической области.

После выбора определенного критерия и уровня значимости все множество возможных значений критерия разбивается на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается, другое — значения, при которых она принимается.

Определение 15. *Критической областью называют множество значений критерия, при которых нулевую гипотезу отвергают.*

Для нахождения критической области достаточно найти ее границы, следовательно, возникает вопрос, как их найти. Находят их исходя из следующего требования:

$$P(\varphi < \varphi_{\text{кр. лев}}) = \alpha/2, \qquad P(\varphi > \varphi_{\text{кр. прав}}) = \alpha/2,$$

где $\varphi_{\text{кр. лев}}$ — левая граница критической области, а $\varphi_{\text{кр. прав}}$ — правая граница критической области.

В случае если распределение случайной величины φ симметрично относительно нуля (как в рассматриваемом примере неполадки робота), то $\varphi_{\text{кр. прав}} = -\varphi_{\text{кр. прав}}$, т. е. достаточно найти одну границу, которую будем обозначать $\varphi_{\text{кр}}$.

С учетом этого условие для нахождения границ критической области можно записать так:

$$P(\varphi < -\varphi_{\text{кр}}) + P(\varphi > \varphi_{\text{кр}}) = 2P(\varphi > \varphi_{\text{кр}}) = \alpha,$$

$$\int_{\varphi_{\text{кр}}}^{+\infty} f_{\varphi}(x)dx = \frac{\alpha}{2},$$

где $f_{\varphi}(x)$ — плотность распределения случайной величины φ.

Для рассматриваемого примера данное нелинейное уравнение приближенно решим на компьютере и получим численное значение для $\varphi_{\text{кр}} = 2.06$.

Этап 5. Расчет наблюдаемого значения и принятие решения. В результате эксперимента случайная величина $\overrightarrow{X_n}$ примет значение $\overrightarrow{x_n}$, а значит, функция $\varphi(\overrightarrow{X_n})$ примет значение $\varphi_{\text{набл}} = \varphi(x_n)$.

Определение 16. *Наблюдаемым значением критерия $\varphi_{\text{набл}}$ называют значение критерия, вычисленное по выборке.*

Если наблюдаемое значение критерия $\varphi_{\text{набл}}$ попадает в критическую область, нулевую гипотезу отвергают как противоречащую экспериментальным данным. Если наблюдаемое значение критерия

попадает в область принятия гипотезы, нет оснований отвергать нулевую гипотезу.

Если в результате 25 наблюдений отклонение средней силы окажется настолько большим, что $|\varphi_{\text{набл}}| > \varphi_{\text{кр}}$, то гипотезу H_0 надо отвергнуть. Если же отклонение средней силы окажется таким, что $|\varphi_{\text{набл}}| \leq \varphi_{\text{кр}}$, то не будет оснований отвергнуть гипотезу H_0.

Пусть в данном примере аналитик получил $\varphi_{\text{набл}} = 3.87$. Это означает, что отклонения статистически значимы. Таким образом, аналитик отвергает нулевую гипотезу о том, что робот откалиброван хорошо, и делает заключение, что робот развивает силу более 600 Н.

3.3. Мощность критерия и некоторые дополнительные замечания

При заданном уровне значимости α можно построить разные критические области. Мы в разобранном в предыдущем пункте примере построили ее симметричной относительно нуля, при этом никак не обосновывали такое решение. Целесообразно ввести в рассмотрение вероятность попадания критерия в критическую область при условии, что основная гипотеза неверна и справедлива конкурирующая гипотеза.

Определение 17. *Мощностью критерия называют вероятность попадания критерия в критическую область при условии, что справедлива конкурирующая гипотеза.*

Другими словами, мощность критерия есть вероятность того, что основная гипотеза будет отвергнута, если верна конкурирующая гипотеза.

Пусть для проверки гипотезы принят определенный уровень значимости и выборка имеет определенный фиксированный объем. Покажем, что строить критическую область надо так, чтобы мощность критерия была максимальной.

Предварительно убедимся, что если вероятность ошибки второго рода (принять неверную гипотезу H_0) равна β, то мощность критерия равна $1-\beta$. Действительно, если β — вероятность ошибки второго рода, т. е. события «принята нулевая гипотеза, причем справедлива конкурирующая», то вероятность противоположного события «отвергнута нулевая гипотеза, причем справедлива конкурирующая», т. е. мощность критерия, равна $1-\beta$.

Пусть мощность $1-\beta$ критерия возрастает, следовательно, уменьшается вероятность β совершить ошибку второго рода, что, конечно, желательно.

Замечание 1. *Поскольку вероятность события «ошибка второго рода допущена» равна β, то вероятность противоположного события «ошибка второго рода не допущена» равна $1-\beta$, т. е. мощности критерия. Отсюда следует, что мощность критерия есть вероятность того, что не будет допущена ошибка второго рода.*

Замечание 2. *Ясно, что чем меньше вероятность ошибок первого и второго рода, тем критическая область «лучше». Однако при заданном объеме выборки уменьшить одновременно и α и β невозможно: если уменьшать α, то β будет возрастать. Например, если принять $\alpha = 0$, то будут приниматься все гипотезы, в том числе и неправильные, т. е. возрастает вероятность ошибки второго рода.*

Ответ на вопрос, как выбрать значение α наилучшим образом, зависит от «тяжести последствий» ошибок для каждой конкретной задачи. Например, если ошибка первого рода повлечет большие потери, а второго — малые, то следует принять, возможно, меньшее значение α. Если α уже выбрано, то, пользуясь теоремой Неймана–Пирсона, можно построить критическую область, для которой β будет минимальным и, следовательно, мощность критерия максимальной.

Замечание 3. *Единственный способ одновременного уменьшения*

вероятности ошибок первого и второго рода состоит в увеличении объема выборки.

3.4. Проверка статистических гипотез в пакетах прикладных программ

В учебниках 70-х гг. описывалась примерно такая последовательность действий при проверке статистических гипотез:

1) рассчитать наблюдаемое значение критерия $\varphi_{\text{набл}}$;
2) задать уровень значимости α, например, равный 0.05, и найти критические точки $\varphi_{\text{кр}}$ в нужной таблице;
3) сравнивая $\varphi_{\text{набл}}$ и $\varphi_{\text{кр}}$, принять решение относительно гипотезы H_0.

При этом необходимо было знать вид критической области (левосторонняя, правосторонняя или двусторонняя), понимать, когда делится пополам уровень значимости, и многое другое. Отягчало ситуацию и то, что таблицы эти оформлялись в каждом справочнике по-разному. Запутаться было несложно.

В современных пакетах прикладных программ проверка статистических гипотез автоматизирована, и подробности этого многоэтапного процесса скрыты от пользователя. Выбирать уровень значимости до начала вычислений не требуется. По введенным данным (выборке) вычисляется $\varphi_{\text{набл}}$, которое пользователю зачастую даже неинтересно, и так называемое p-value — вероятность получить такое или большее отклонение при условии справедливости нулевой гипотезы. Именно на основе p-value принимается решение. Пользователь устанавливает (исходя из внестатистических соображений) некоторый уровень значимости α, например, 0.05 или 0.01. Если в результате расчетов p-value $< \alpha$, то нулевую гипотезу надо отвергнуть как противоречащую экспериментальным данным.

В рассмотренном примере про робота p-value $= 0.004 < 0.01$, следовательно, можно считать, что на уровне значимости $\alpha = 0.01$ нулевую гипотезу надо отвергнуть.

3.5. Заключительные замечания

1. Часто приходится слышать: «Если уровень значимости меньше 0.05, то нулевую гипотезу отвергаем». Это высказывание лишено смысла, так как уровень значимости — это величина, которая фиксируется самим исследователем, а не вычисляется на основе выборочных данных.

2. Критические области бывают одно- и двусторонними.

3. Статистические гипотезы бывают простыми и сложными. Статистическая гипотеза, однозначно определяющая распределение генеральной совокупности, называется простой. Статистическая гипотеза, утверждающая принадлежность распределения к некоторому семейству распределений, называется сложной.

Задания для самостоятельной работы

1. В качестве примера возьмем вымышленное государство, в котором проводят реформу. Известно, что в прошлом году расходы населения на развлечения (кино, кафе, катание на горных лыжах и т. д.) распределялись по нормальному закону и составляли в среднем 700 руб. в месяц на одного человека. По заявлению министра труда и социального развития в этом году в связи с экономическим ростом граждане стали богаче и на развлечения смогли тратить больше. Министр заявляет, что причина этого в реализуемых мерах по развитию экономики. Однако некоторые эксперты утверждают, что это всего лишь домыслы, и граждане не стали лучше жить.

Требуется понять, выросли ли траты на развлечения в этом году. 25 респондентам был задан вопрос: «Сколько денег на развлечения вы потратили в этом месяце?» Результаты представлены в таблице. Проверьте, согласуются ли фактические данные с заявлением министра.

678	698	704	716	728
724	698	709	692	688
734	705	690	728	732
741	685	741	737	754
684	727	712	706	720

Какую критическую область нужно рассмотреть, одно- или двустороннюю? Нулевая гипотеза будет простой или сложной? А альтернативная?

2. Что такое мощность критерия? Как меняется мощность критерия при изменении уровня значимости?

3. При проверке статистических гипотез назначается уровень значимости. Уровнем значимости называется вероятность ошибки первого рода, т. е. вероятность ошибочно отклонить верную нулевую гипотезу. Студент Василий не изучал статистику и рекомендует взять уровень значимости равным нулю, чтобы ошибок при проверке гипотез не было вовсе. К каким негативным последствиям приведет такой выбор?

4. В учебном пособии по спортивной метрологии В. В. Афанасьева[5] находим следующий вопрос для самоконтроля: «Укажите, при каком уровне значимости принимается гипотеза H_0». И далее предложены ответы:

1) $\alpha \geq 0.05$; 2) $\alpha \leq 0.05$; 3) $\alpha \geq 0.5$; 4) $\alpha \leq 0.5$.

В ответах находим: «Если $\alpha \geq 0.05$, то принимается гипотеза H_0».

Объясните, почему этот вопрос некорректно сформулирован и, по сути, лишен смысла. Как его исправить? 5. Проведено три эксперимента при уровне значимости $\alpha = 0.05$:

- в первом эксперименте найдено p-value $= 0.02$, нулевая гипотеза H_0 отвергнута;

- во втором эксперименте найдено p-value $= 0.049$, нулевая гипотеза H_0 отвергнута;

- в третьем эксперименте найдено p-value $= 0.15$, нулевая гипотеза H_0 принята.

В каком из этих трех случаев вероятность верного решения наибольшая?

[5]Спортивная метрология : учебник для среднего проф. образования / В. В. Афанасьев, И. А. Осетров, А. В. Муравьев, П. В. Михайлов ; отв. ред. В. В. Афанасьев. 2-е изд., испр. и доп. М. : Юрайт, 2017.

4. Корреляционный анализ

В этой и последующих главах мы будем рассматривать те или иные аспекты зависимостей между двумя или большим числом величин. При этом будем опираться на общую теорию проверки статистических гипотез. Так как нам предстоит исследовать очень обширную тему, то полезно начать с общего обзора.

4.1. Введение и основные идеи

При изложении материала данного параграфа мы опирались на работу [6]. Большая часть работ по этой теме возникла в связи с задачей о совместном распределении пары случайных величин; ее можно назвать задачей о статистической зависимости. Существует иная область математики, касающаяся зависимостей строго функционального вида между величинами (как, например, зависимости в классической физике). Указанный вид зависимостей тоже представляет статистический интерес, потому что функционально связанные величины подвержены ошибкам наблюдений или измерений. Назовем это задачей о *функциональной зависимости*. В рамках настоящего пособия мы будем заниматься только задачей о *статистической зависимости*, в которой величины (кроме вырожденных случаев) не связаны функционально и, кроме того, могут быть подвержены ошибкам наблюдений и измерений. Мы будем рассматривать их просто как совокупность случайных величин, подчиненных некоторому совместному распределению[6].

В самой области статистической зависимости необходимо провести разграничение. Нас может интересовать либо *взаимозависимость* между несколькими величинами (не обязательно между всеми), либо *зависимость* одной или большего числа величин от остальных. Например, можно рассмотреть вопрос, существует ли связь

[6]Ранее мы рассмотрели понятие случайной величины. О том, как определяется совокупность случайных величин, можно прочитать в работе [1, гл. 14.]

между доходом семьи и расходами на роскошь (дорогие машины, ювелирные украшения и т. п.); при такой постановке это есть задача о взаимозависимости. Но если мы хотим, используя измерения дохода семьи, получить информацию об ожидаемых тратах на роскошь, то мы приходим к задаче о зависимости расходов от доходов. Это пример ситуации, в которой может представлять интерес как взаимозависимость, так и зависимость. С другой стороны, имеются ситуации, в которых интересна только зависимость. Связь между величиной урожая и количеством выпавших осадков представляет собой пример существенной асимметрии. Здесь из внестатических соображений понятно, что дожди влияют на урожай, а урожай не воздействует на дожди. Таким образом, мы должны изучать зависимость урожая от дождей.

Прежде чем перейти к изложению теории корреляции, которую в конце XIX — начале XX в. развивали Пирсон и Юл, сделаем одно общее замечание.

Замечание 4. *Статистическая зависимость, какой бы сильной она ни была, никогда не может доказать причинную связь: наши идеи о причине возникают вне статистики, в конечном счете — из некоторой другой теории.*

Даже в простом примере о величине урожая и количестве осадков мы не имеем статистических причин для отказа от идеи, что дожди зависят от урожая: отказ сделан на основе совершенно других соображений. И даже если бы дожди и урожай были в полном функциональном соответствии, то мы все равно не подумали бы провозгласить эту «очевидную» причинную связь. Нет необходимости углубляться в философское обсуждение данного вопроса; нужно лишь еще раз обратить внимание на то, что любая статистическая зависимость логически не влечет причинной.

Бернард Шоу блестяще сказал об этом в своем предисловии к «Доктору на распутье»: «Даже опытные статистики часто оказы-

ваются не в состоянии оценить, до какой степени смысл статистических данных искажается молчаливыми предположениями их интерпретаторов... Легко доказать, что ношение цилиндров и зонтиков расширяет грудную клетку, удлиняет жизнь и дает относительный иммунитет от болезней... Университетский диплом, ежедневная ванна, обладание тридцатью парами брюк, знание музыки Вагнера, скамья в церкви, — короче, все, что подразумевает большие средства и хорошее воспитание... может быть с помощью статистики представлено как магические чары, дарующие привилегии любого сорта. Математик, чьи корреляции привели бы в восхищение Ньютона, может, собирая данные и делая из них выводы, впасть в совершенно грубые ошибки на основе таких же популярных заблуждений, как описанные выше».

Хотя Шоу в данном случае отстаивает сомнительную точку зрения, его логика обоснована. Последователи Пирсона и Юла в первом приступе энтузиазма, порожденного корреляционной техникой, легко делали опрометчивые выводы. Это продолжалось до тех пор, пока (спустя двадцать лет после написанного Шоу) Юл не напугал статистиков примерами высоких корреляций, которые, очевидно, не выражали причинных связей: например, количество самоубийств за год было сильно коррелировано с принадлежностью к англиканской церкви. Большинство этих «бессмысленных» корреляций действует через сопутствующие изменения во времени. Упомянутые примеры имели благотворный эффект, доводя до сознания статистиков, что причинная зависимость не может быть выведена ни из какого наблюдаемого совместного изменения, даже самого тесного. Впоследствии статистики впали в другую крайность: в 70-х гг. XX в. корреляционный анализ стал совершенно немодным. Имеется, однако, широкое поле приложений (например, социальные науки и психология), где характеры причин еще недостаточно хорошо поняты для того, чтобы корреляционный анализ был заменен более специфическими «структурными» статистическими методами. Есть, кроме того, обширная

область многомерного анализа, где вычисление и исследование матрицы коэффициентов корреляции является необходимой прелюдией к детальному статистическому анализу. Все это делает необходимым изучение корреляционного анализа, основы которого приведены в этой главе.

4.2. Линейная и нелинейная зависимость, корреляция

Пусть даны наблюдения за двумя случайными величинами X и Y, которые мы будем трактовать как цену на нефть в долларах за баррель и доход в бюджет в миллионах рублей. Требуется проверить, существует ли статистическая зависимость между этими величинами, и если да, то какая и насколько тесная. Для ответа на этот вопрос используем корреляционный анализ.

X	30	69	86	56	44	97	53
Y	204	267	313	207	130	320	237

X	66	39	29	34	31	92
Y	260	172	117	196	169	262

Определение 18. *Корреляционный анализ — метод обработки статистических данных, с помощью которого измеряется теснота связи между двумя или более переменными.*

Исследование стоит начать с графического изображения данных на плоскости XOY. На рис. 2 видно, что точки образовали некоторое облако и выстроились вдоль воображаемой наклонной прямой, т. е. подчинены *линейной статистической связи*. Однако данное наблюдение должно быть проверено, и в этом нам поможет разобранная в предыдущей главе общая теория проверки статистических гипотез.

Для того чтобы пояснить значение термина «линейный», вспомним общий вид уравнения прямой: $y = kx + b$. Здесь y — зависимая

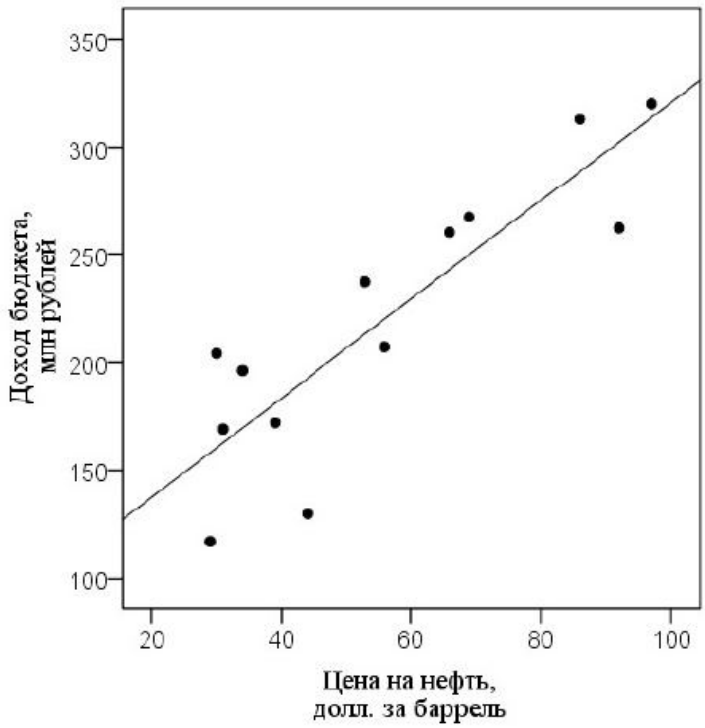


Рис. 2. Зависимость дохода бюджета, млн руб., от цен на нефть

величина; x — независимая величина; k — угловой коэффициент; b — свободный член. Содержательный смысл этих величин следующий: если $x = 0$, то $y = b$, т. е. b — это начальное смещение; при увеличении x на единицу y увеличивается на k единиц.

Связи между случайными величинами можно разделить на линейные и нелинейные (рис. 3). Линейные статистические связи — это такие статистические связи, которые хорошо описываются уравнением прямой. В рассматриваемом примере можно сказать, что точки на плоскости располагаются вдоль прямой. Нелинейные статистические связи — это такие статистические связи, которые плохо описываются уравнением прямой или не описываются вовсе. Примером нелинейной связи может служить зависимость скорости роста микроорганизмов в среде от концентрации сахара: сначала рост концентрации сахара способствует увеличению количества микроорганизмов, но с

определенного уровня сахар начинает действовать как консервант и останавливает рост микроорганизмов.

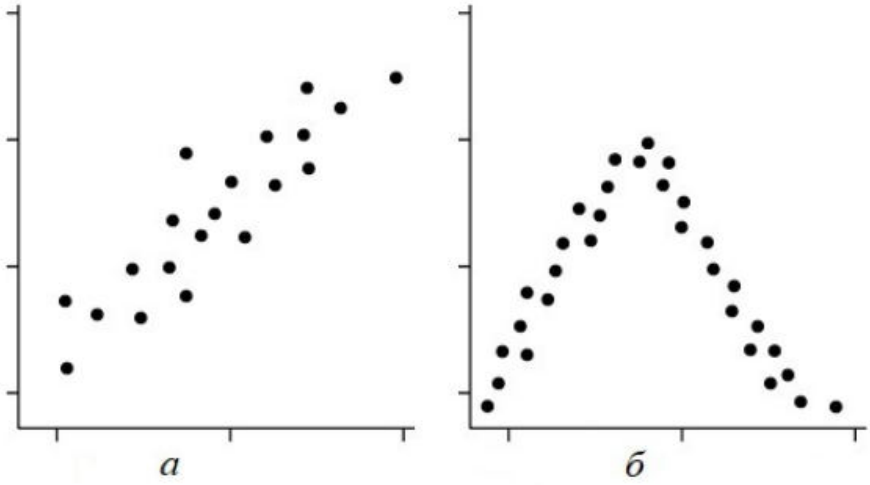


Рис. 3. Линейная и нелинейная зависимость: *а* — линейная связь; *б* — нелинейная связь

Для анализа нелинейных связей может применяться корреляционное отношение, определяемое через отношение межгрупповой дисперсии к общей:

$$\eta^2_{Y|X} = 1 - M\left[\frac{D(Y|X)}{D(Y)}\right],$$

где $D(Y)$ — дисперсия Y; $D(Y|X)$ — условная дисперсия Y при данном X, характеризующая рассеяние Y около условного математического ожидания $M(Y|X)$ при данном значении X. Мы в дальнейшем будем анализировать только линейные связи.

Определение 19. *Корреляция — это линейная связь между парой случайных величин.*

Две случайные величины могут быть связаны более тесной линейной связью, тогда соответствующее корреляционное облако будет узкое, или менее тесной линейной связью, тогда соответствующее корреляционное облако будет широкое (рис. 4).

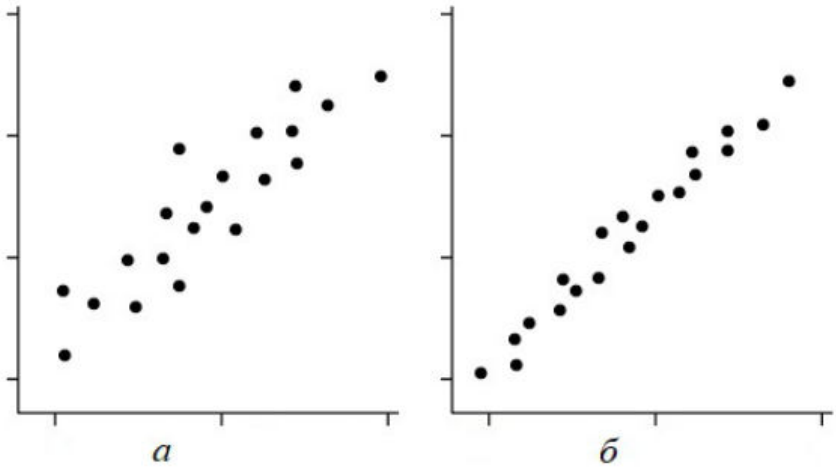


Рис. 4. Теснота линейной зависимости: *а* — нетесная линейная связь; *б* — тесная линейная связь

Для оценки тесноты линейной связи между парой случайных величин введем понятие коэффициента корреляции[7].

Определение 20. *Коэффициент корреляции — это числовая величина, характеризующая тесноту линейной связи между парой случайных величин.*

Коэффициент корреляции случайных величин X и Y обозначают r_{XY}. Найти коэффициент корреляции можно по следующей формуле:

$$r_{XY} = \frac{M(XY) - M(X)M(Y)}{\sigma_X \sigma_Y},$$

где $M(X)$ — математическое ожидание случайной величины X; σ_X — среднее квадратическое отклонение случайной величины X; $M(XY)$ — математическое ожидание произведения случайных величин X и Y.

Перечислим свойства коэффициента корреляции:

[7]Здесь рассматривается коэффициент корреляции Пирсона, который находят для пары случайных величин в интервальной шкале. Коэффициенты корреляции Спирмена и Кендалла будут рассмотрены далее.

1. Коэффициент корреляции не превосходит единицы по модулю, $-1 \leq r_{XY} \leq 1$.
2. Если величины связаны строгой функциональной связью и с увеличением X величина Y также возрастает, то $r_{XY} = 1$. В этом случае точки корреляционного облака идеально ложатся на прямую, направленную вверх.
3. Если величины связаны строгой функциональной связью и с увеличением X величина Y убывает, то $r_{XY} = -1$. В этом случае точки корреляционного облака идеально ложатся на прямую, направленную вниз.
4. Если $r_{XY} = 0$, то величины некоррелированы. В этом случае точки корреляционного облака вдоль прямой не ложатся и обычно хаотично разбросаны.

Поясним связь между понятиями «зависимость» и «коррелированность». Поскольку коррелированность — это частный случай зависимости (линейная зависимость), то из коррелированности следует зависимость. Обратное неверно: из зависимости не следует коррелированность (рис. 3, *б*, и рис. 5, нижняя строка).

Замечание 5. *Часто, желая сделать свою речь более наукообразной и показать свою «образованность», люди допускают ошибку, говоря «это некоррелированные величины», имея в виду то, что величины несвязаны. Обращаем внимание читателей, что зависимость и коррелированность — это не синонимы.*

Если величины связаны линейной связью, то коэффициент корреляции показывает только тесноту корреляционного облака, отсутствие «зашумленности», но не показывает наклон прямой (рис. 5, вторая строка). То есть коэффициент корреляции не показывает, на сколько изменяется одна величина при изменении другой на единицу. Если такая оценка требуется, то необходимо провести регрессионный анализ.

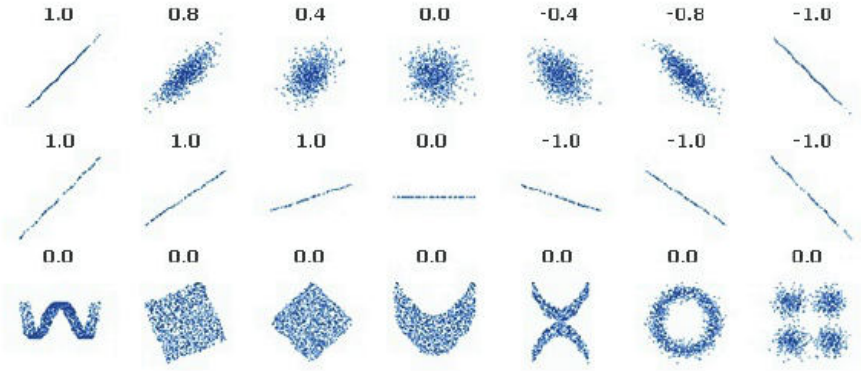


Рис. 5. Виды корреляционных облаков

Вернемся к поставленной задаче о связи цен на нефть и дохода бюджета. Здесь мы работаем с выборкой, которая отражает (как мы надеемся, правильно) пропорции генеральной совокупности. В данном случае генеральная совокупность является двумерной случайной величиной, первая компонента которой — это цена на нефть, вторая — доход бюджета. Для оценки коэффициента корреляции находят выборочный коэффициент корреляции r^*_{XY}. Мы будем проводить расчеты только на компьютере с использованием статистических пакетов, где формулы для нахождения r^*_{XY} заранее запрограммированы.

Пусть, например, мы получили значение $r^*_{XY} \approx 0.21$. Означает ли это, что величины коррелированы? Вообще говоря, нет. Возможно, что это положительное значение — результат случайности, и по другой выборке мы получим совсем иное, возможно, отрицательное значение или очень близкое к нулю. Для уверенного ответа на вопрос о наличии линейной связи между величинами надо следовать нижеописанной процедуре.

Этап 1. Формулировка основной и альтернативных гипотез. Гипотеза H_0 состоит в том, что $r_{XY} = 0$, а полученное отличие расчетного значения r^*_{XY} от нуля — результат случайности, гипотеза H_1 состоит в том, что величины коррелированы, т. е. $r_{XY} \neq 0$.

Этап 2. Назначение уровня значимости α. В наших задачах мы, следуя общепринятым в социологии, психологии и иных гуманитарных науках рекомендациям, положим $\alpha = 0.05$.

Этап 3. Нахождение величины[8] p-value. Если p-value $< \alpha$, то нулевую гипотезу отвергают как противоречащую экспериментальным данным, иначе нет оснований отвергнуть нулевую гипотезу.

В рассматриваемом примере получаем p-value $= 9.35 \cdot 10^{-5}$. Таким образом, нулевую гипотезу отвергаем. Содержательный смысл величины p-value следующий: p-value — это вероятность получить такую (или более тесно связанную, с более узким корреляционным облаком) выборку, если бы величины действительно были бы некоррелированы.

Этап 4. Нахождение коэффициента корреляции, если нулевая гипотеза была отвергнута. Только если на этапе 3 было показано, что коэффициент корреляции статистически значимо отличается от нуля, находят его числовое значение. В данном примере получено $r_{XY} = 0.874$. Можно сказать, что доходы бюджета тесно связаны с ценами на нефть.

4.3. Ранговая корреляция

Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные. Для переменных, представленных в интервальной шкале, необходимо использовать коэффициент корреляции Пирсона. Если по меньшей мере одна из двух переменных имеет порядковую шкалу либо не является нормально распределенной, то необходимо использовать ранговую корреляцию Спирмена или Кендалла.

Так, например, для выявления связи между степенью артериальной гипертензии и функциональным классом хронической сердечной недостаточности следует вычислять ранговые коэффициенты корреляции Спирмена или Кендалла.

[8] В некоторых статистических пакетах эта величина называется Sig.

4.4. Ложная корреляция

Существуют примеры весьма значимых и высоких корреляций между совершенно не связанными друг с другом величинами. Например, есть связь между объемами вырабатываемой электроэнергии в США и длиной прыжков на чемпионатах мира. Однако ясно, что эти величины не могут быть связанными. Причина корреляции состоит в том, что в этой системе неявно присутствует третья величина — время. В течение XX в. объемы выработки электроэнергии только увеличивались. Точно так же совершенствовались обувь спортсменов, покрытие стадионов, качество тренировок, в результате спортсмены стали прыгать дальше. Если исключить влияние времени, то связи между переменными не будет.

Определение 21. *Ложная корреляция — корреляция, которая возникла не в результате прямого соотношения между оцениваемыми переменными, а в результате их связей с третьей переменной (или четвертой, или более); при этом нет никакой связи, объединяющей эти переменные.*

Для исключения влияния третьей переменной вычисляют частный коэффициент корреляции при исключенном влиянии третьей.

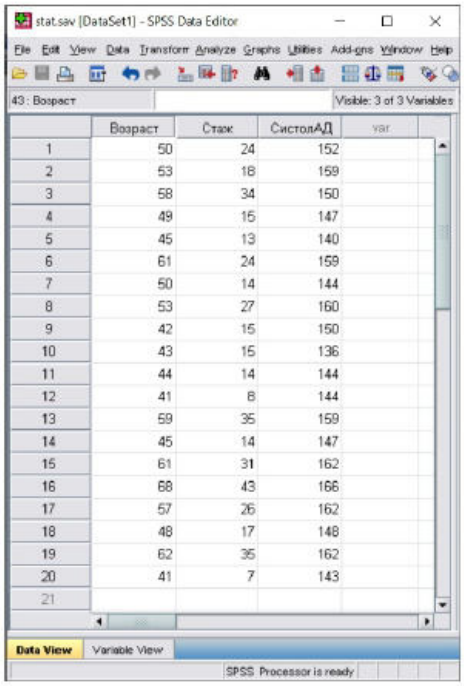
4.5. Использование пакетов программ для проведения корреляционного анализа

Для расчета коэффициента корреляции в пакете SPSS выбирают последовательно пункты меню `Analyze - Correlate - Bivariate`, заносят две анализируемые переменные в область `Variables`. В зависимости от того, что надо найти, используют коэффициент Пирсона, Кендалла или Спирмена. Нажимают кнопку OK.

Для расчета частного коэффициента корреляции двух величин X и Y при исключенном влиянии величины Z в пакете SPSS выбирают последовательно пункты меню `Analyze - Correlate - Partial`. Заносят две анализируемые переменные X и Y в область `Variables`,

а переменную Z — в область `Controlling for`. Нажимают кнопку OK.

П р и м е р. Пусть в моногороде[9] проведено поперечное исследование работников градообразующего предприятия с целью выявления связи между наличием (а если есть, то между стадией) артериальной гипертензии и вредным стажем рабочих. Исследователи — санитарные врачи — предположили, что вредный стаж на данном производстве является причиной развития сердечно-сосудистой патологии, в частности артериальной гипертензии. Данные были занесены в таблицу (рис. 6).



	Возраст	Стаж	СистолАД	var
1	50	24	152	
2	53	18	159	
3	58	34	150	
4	49	15	147	
5	45	13	140	
6	61	24	159	
7	50	14	144	
8	53	27	160	
9	42	15	150	
10	43	15	136	
11	44	14	144	
12	41	8	144	
13	59	35	159	
14	45	14	147	
15	61	31	162	
16	68	43	166	
17	57	26	162	
18	48	17	148	
19	62	35	162	
20	41	7	143	
21				

Рис. 6. Данные медосмотра рабочих: возраст, лет; стаж, лет; систолическое артериальное давление, мм. рт. ст.

Если просто найти коэффициент корреляции Пирсона для ста-

[9]Моногород — населенный пункт, основанный при градообразующем предприятии с целью обеспечения производства трудовыми ресурсами.

жа и САД, то окажется, что между величинами существует тесная связь: $r_{\text{стаж, САД}} = 0.805$, $p < 0.001$, которая может быть неправильно интерпретирована как производственная обусловленность артериальной гипертензии у данных рабочих (рис. 7). Почему мы говорим «неправильно интерпретирована»?

Correlations

		Возраст	Стаж	СистолАД
Возраст	Pearson Correlation	1	,924	,861
	Sig. (2-tailed)		,000	,000
	N	20	20	20
Стаж	Pearson Correlation	,924	1	,805
	Sig. (2-tailed)	,000		,000
	N	20	20	20
СистолАД	Pearson Correlation	,861	,805	1
	Sig. (2-tailed)	,000	,000	
	N	20	20	20

а

Correlations

Control Variables			Стаж	СистолАД
Возраст	Стаж	Correlation	1,000	,047
		Significance (2-tailed)	.	,849
		df	0	17
	СистолАД	Correlation	,047	1,000
		Significance (2-tailed)	,849	.
		df	17	0

б

Рис. 7. Анализ связей между возрастом, стажем и уровнем систолического артериального давления у рабочих на медосмотре с помощью коэффициента корреляции: *а* — коэффициент корреляции Пирсона; *б* — частный коэффициент корреляции стажа и систолического артериального давления при исключенном влиянии возраста

Дело в том, что люди, проработавшие на производстве более 40 лет, уже немолоды, им более 60 лет, и артериальная гипертония у них могла развиться в силу их возраста, как и у представителей популяции в целом, т. е. не является производственно обусловленной.

В моногородах связь возраста и стажа особенно сильная:

$r_{\text{возраст, стаж}} = 0.924$, $p < 0.001$, а потому можно предположить, что у людей, проработавших на производстве более 30 лет, артериальная гипертензия возникла не в связи с вредным производством, а в силу их возраста. Найдем частный коэффициент корреляции стажа и САД при исключенном влиянии возраста: $r_{\text{стаж, САД / возраст}} = 0.047$, $p = 0.849$, он недостоверно отличается от нуля. Это показывает, что имеет место ложная корреляция стажа и САД.

Задания для самостоятельной работы

1. Известно, что две величины связаны корреляционной связью, следует ли отсюда, что они связаны и причинно-следственной связью?

2. Рассматривая пожары в конкретном городе, можно выявить весьма высокую корреляцию между ущербом, который нанес пожар, и количеством пожарных, участвовавших в ликвидации пожара, причем эта корреляция будет положительной. Следует ли отсюда вывод, что увеличение количества пожарных приводит к увеличению причиненного ущерба? Можно ли для минимизации ущерба от пожаров ликвидировать пожарные бригады?

3. В 1965 г. английский эпидемиолог сэр Остин Брэдфорд Хилл сформулировал критерии причинно-следственной связи. Критерии Хилла представляют собой список из девяти принципов, которые могут быть полезны при доказательстве причинно-следственной связи между явлениями.

Ознакомьтесь со статьей: *Hill A. B.* The Environment and Disease: Association or Causation? // Proceedings of the Royal Society of Medicine. 1965. Vol. 58, № 5. P. 295–300.

4. Известно, что две величины связаны функциональной связью, следует ли отсюда, что они связаны и корреляционной связью? Приведите пример.

5. Профессор Р. А. Шамойлова учит студентов проводить корреляционный анализ. На лекции она дала следующее определение: «Корреляция — это статистическая зависимость между случайными величинами,

не имеющая строгого функционального характера, при которой изменение одной из случайных величин приводит к изменению математического ожидания другой». Согласны ли вы с профессором?

Также Р. А. Шамойлова утверждает, что для расчета коэффициента корреляции Пирсона необходимо, чтобы совокупность значений всех факторных и результативных признаков подчинялась многомерному нормальному распределению. Не ошибается ли она?

Далее профессор утверждает, что при проведении корреляционного анализа первым делом выборку нужно проверить на нормальность, и в случае если объем выборки недостаточен для проведения формального тестирования, то закон распределения определяется визуально на основе корреляционного поля. Если в расположении точек на этом поле наблюдается линейная тенденция, то исходные данные подчиняются нормальному закону распределения. Права ли профессор на этот раз?

6. В учебном пособии по спортивной метрологии В. В. Афанасьева[10] находим следующее определение: «Коэффициент корреляции — это статистический показатель зависимости двух случайных величин. Коэффициент корреляции может принимать значения от -1 до $+1$. При этом значение -1 будет говорить об отсутствии корреляции между величинами, 0 — о нулевой корреляции, а $+1$ — о полной корреляции величин. То есть чем ближе значение коэффициента корреляции к $+1$, тем сильнее связь между двумя случайными величинами».

Объясните, почему это определение неверное. Какие еще здесь ошибки?

7. В таблице приведены наблюдения за ценами на кормовое зерно и розничными ценами на молоко. Коррелированы ли эти величины на уровне значимости 0.05? Рассмотреть двустороннюю критическую область. Если да, то чему равен коэффициент корреляции?

Зерно	21	24	26	22	21	26	23
Молоко	16	26	25	26	25	28	19

[10]Спортивная метрология : учебник для среднего проф. образования.

Зерно	24	20	21	21	20	26	27
Молоко	22	29	18	19	24	30	34

Ответы и указания

6. «Коэффициент корреляции – это статистический показатель *зависимости* двух случайных величин». Как известно, коэффициент корреляции – это мера *линейной связи*, он не является мерой зависимости в общем случае. То есть величины могут быть зависимыми, но некоррелированными, так как их зависимость носит нелинейный характер. Данная неточность (а на самом деле принципиальная ошибка) очень распространена даже среди людей, которые позволяют себе писать книги по статистике.

Значение -1, равно как и $+1$, говорит о линейной связи между величинами. Таким образом, утверждение о том, что «значение -1 будет говорить об отсутствии корреляции между величинами» является ошибочным. Далее авторы учебника разделяют понятия «нулевая корреляция» и «отсутствие корреляции», хотя на самом деле это синонимы. Авторы используют такую характеристику корреляции, как «полная». Такой характеристики нет.

Неверной является и заключительная фраза о том, что «чем ближе значение коэффициента корреляции к $+1$, тем сильнее связь между двумя случайными величинами». Нужно говорить, чем значение коэффициента корреляции *по модулю* ближе к 1, тем сильнее (а лучше сказать, теснее) связь между двумя случайными величинами.

5. Регрессионный анализ

После того как методами корреляционного анализа установлено, что существует линейная связь между признаками, естественно возникает вопрос, как описать эту связь в виде формулы. Возвратимся к примеру из предыдущей главы: пусть установлено, что цены на нефть влияют на доходную часть бюджета. Требуется узнать:

1) на сколько увеличивается доход бюджета при увеличении цены на нефть на один доллар за баррель;
2) какой ожидается доход бюджета, если цена на нефть установится на уровне 80 долл. за баррель.

Материал излагается в следующей последовательности. В п. 5.1 в рамках теоретико-вероятностного подхода мы рассматриваем систему линейно зависимых случайных величин Y и X, распределения которых нам известны, и описываем связь между ними в виде уравнения линейной регрессии — по сути, находим условное математическое ожидание Y по X.

В п. 5.2 мы продолжаем рассматривать систему линейно зависимых случайных величин Y и X, но на этот раз распределения Y и X неизвестны, есть лишь набор из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, за X и Y. По этому набору строим выборочное уравнение линейной регрессии.

В п. 5.4 мы рассматриваем множественную линейную регрессию и обсуждаем, какие новые проблемы возникают при увеличении числа предикторов.

В п. 5.5 ситуация усложняется — мы рассматриваем величины, связь которых близка к линейной, но таковой не является, например, $Y = X^{1.2} + \varepsilon$. Ситуация осложняется тем, что точный вид связи нам априори неизвестен. По набору из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, строим выборочное уравнение линейной регрессии, которая лишь упрощенно описывает нелинейную связь.

5.1. Уравнение линейной регрессии. Теоретико-вероятностный подход

В этом параграфе мы следуем теоретико-вероятностному подходу. Мы не работаем с выборками и не делаем оценок. Рассматривается система случайных величин, описанных функцией или плотностью совместного распределения. Затем на основе этих данных строится функция линейной регрессии Y по X.

Пусть $f_{X,Y}(x,y)$ — плотность совместного распределения случайных величин X и Y. Тогда маргинальная плотность распределения случайных величин X и Y определяется следующим образом:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy, \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx.$$

Для каждого фиксированного значения x случайной величины X и значения y случайной величины Y условные распределения Y по X и X по Y соответственно определяются по формулам:

$$f_Y(y|X=x) = \frac{f_{X,Y}(x,y)}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy}, \quad f_X(x|Y=y) = \frac{f_{X,Y}(x,y)}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx}.$$

Далее определяются условные математические ожидания Y при фиксированном x и X при фиксированном y:

$$M[Y|X=x] = \frac{\int\limits_{-\infty}^{+\infty} y\, f_{X,Y}(x,y)\,dy}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy}, \quad M[X|Y=y] = \frac{\int\limits_{-\infty}^{+\infty} x\, f_{X,Y}(x,y)\,dx}{\int\limits_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx}.$$

Приведенные соотношения соответственно определяют регрессии[11] Y по X и X по Y (кривые регрессии). Первое из них выражает зависимость среднего значения величины Y от x. Данная за-

[11] Далее рассмотрим только регрессию Y по X, так как для регрессии X по Y все аналогично.

висимость, вообще говоря, нелинейная, является функциональной, а не статистической.

Особый интерес представляют регрессии, которые выражаются линейной функцией. В качестве примера можно рассмотреть многомерное нормальное распределение вектора $\mathbf{X}$ с математическим ожиданием $\mathbf{m} \in \mathbb{R}^n$ и ковариационной матрицей Σ размерности $n \times n$ с плотностью распределения

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{x}-\mathbf{m})}, \quad \mathbf{x} \in \mathbb{R}^n,$$

где $|\Sigma|$ — определитель матрицы Σ; Σ^{-1} — матрица, обратная к Σ.

При $n = 2$ плотность двумерного невырожденного (коэффициент корреляции r по модулю не равен единице) нормального распределения можно записать в виде:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times$$
$$\times \exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2} - r\frac{2(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]\right\}.$$

Оказывается, что условное математическое ожидание X_2 имеет вид

$$M[X_2|X_1 = x_1] = m_2 + r\frac{\sigma_2}{\sigma_1}(x_1 - m_1),$$

т. е. выражается линейной функцией. На практике это имеет огромное значение, так как многомерное нормальное распределение часто встречается в силу центральной предельной теоремы, а линейные функции регрессии очень удобны в применении и легко интерпретируются.

Для пары случайных величин X и Y функция регрессии имеет линейный вид не только в случае, если они подчиняются двумерному нормальному распределению. Теорема о необходимом и достаточном условии линейности регрессии есть, например, в [6, с. 467].

Для упрощения математической стороны изложения мы нало-

жим ограничения[12] на случайные величины X и Y. Пусть условное распределение Y относительно своего среднего (которое, как и раньше, является функцией от x) одно и то же для любого x, т. е. только среднее значение Y изменяется при изменении x. Говорят, что Y имеет «однородные ошибки» [6, с. 468]. Таким образом, существует случайная величина ε такая, что

$$Y = M[Y|X = x] + \varepsilon.$$

В частности, когда регрессия линейная, имеем

$$Y = \beta_1 X + \beta_0 + \varepsilon.$$

Поясним сказанное на примере. Пусть есть две случайные величины X и Y, которые можно трактовать, например, как объем бензина в баке автомобиля в момент начала поездки и максимальное расстояние, которое можно проехать без дозаправки по равнинной местности[13]. Из внестатистических соображений известно, что каждый дополнительный литр бензина в баке увеличивает дальность поездки на 8 – 10 км в зависимости от машины. Соответственно, можно с высокой точностью считать, что связь между X и Y линейная.

Очевидно, однако, что мы не можем написать равенство $Y = \beta_1 X + \beta_0$, где β_0 и β_1 — числа, так как при одной и той же заполненности бака дальность поездки может быть различной. Причина заключается в воздействии прочих факторов: ветер, дождь, стиль вождения и т. д. Все эти прочие факторы: 1) в меньшей степени влияют на дальность поездки; 2) не могут быть учтены заранее, а потому мы их в совокупности будем считать помехами и для их учета введем случайную величину ε такую, что $M[\varepsilon] = 0$.

[12] Эти ограничения несильно обременительны и в прикладных задачах обычно выполняются хотя бы приближенно.

[13] В будущем, говоря о дальности поездки, слово «максимальная» будем подразумевать.

Замечание 6. *Для упрощения изложения будем предполагать, что ε распределено нормально. Несложно проверить, что от этого требования можно отказаться, при этом формулы для оценки коэффициентов β_0 и β_1 линейной регрессии останутся справедливы.*

Теперь можно записать такое равенство:

$$Y = \beta_1 X + \beta_0 + \varepsilon.$$

Поскольку в прикладных задачах априори известен только вид уравнения, а конкретные значения коэффициентов β_0 и β_1 неизвестны, естественно выбрать их так, чтобы, зная значение, которое в эксперименте приняла величина X, наиболее точно спрогнозировать значение, которое примет величина Y.

Таким образом, мы приходим к постановке задачи построения линейной регрессии:

$$M[Y - \beta_1 X - \beta_0]^2 \xrightarrow[\beta_0,\,\beta_1]{} \min.$$

Теорема. *Линейная среднеквадратическая регрессия Y на X имеет вид:*

$$f(x) = m_Y + r\frac{\sigma_Y}{\sigma_X}(x - m_X),$$

где m_Y, m_X, σ_Y, σ_X — математические ожидания и средние квадратические отклонения случайных величин Y и X соответственно; r — коэффициент корреляции случайных величин Y и X.

Доказательство. Рассмотрим функцию $F(\beta_0, \beta_1) = M[Y - \beta_1 X - \beta_0]^2$.

Пользуясь формулами $M[X^2] = M^2[X] + D[x]$ и $M[XY] = M[X]M[Y] + \operatorname{cov}(X, Y) = M[X]M[Y] + r\sigma_X\sigma_Y$, получим

$$F(\beta_0, \beta_1) = \sigma_Y^2 + \beta_1^2\sigma_X^2 - 2r\sigma_X\sigma_Y\beta_1 + (m_Y - \beta_1 m_X - \beta_0)^2.$$

Исследуем функцию F на минимум, приравняв частные производные к нулю:

$$\frac{\partial F}{\partial \beta_0} = -2(m_Y - \beta_1 m_X - \beta_0) = 0,$$

$$\frac{\partial F}{\partial \beta_1} = 2\beta_1 \sigma_X^2 - 2r\sigma_X \sigma_Y = 0.$$

Откуда получаем

$$\beta_0 = m_Y - r\frac{\sigma_Y}{\sigma_X} m_X, \qquad \beta_1 = r\frac{\sigma_Y}{\sigma_X}.$$

Здесь, пользуясь тем, что нам известны распределения X и Y, мы выводим оптимальные значения коэффициентов β_0, β_1. Это не оценки, а точные значения, полученные в результате минимизации остаточной дисперсии.

Сделаем небольшое замечание по терминологии. Функцию $f\colon \mathbb{R} \to \mathbb{R}$ называют функцией регрессии. Аргументом и значением функции регрессии являются числа, а не случайные величины. Если функция f имеет вид $\beta_1 x + \beta_0$, то регрессию называют (парной) линейной; случай множественной линейной регрессии аналогичен. Независимые переменные иначе называют предикторами, регрессорами или факторами. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

5.2. Выборочное уравнение линейной регрессии. Случай линейной связи

Аналитическая теория регрессии, представленная в п. 5.1, требует точного знания функции распределения рассматриваемой системы случайных величин, а потому интересна для теории вероятностей, но не для прикладной статистики.

Мы продолжаем рассматривать систему линейно зависимых случайных величин Y и X, но на этот раз распределения Y и X неизвестны, есть лишь набор из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, за X и Y. Предполагается, что не все x_i равны между собой.

Линейная модель, в рамках которой мы работаем, имеет вид:

$$y_i = b_1 x_i + b_0 + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$

В классической линейной регрессии предполагается, что выполнены следующие условия:

1) факторы и случайные ошибки — независимые случайные величины;
2) случайные ошибки модели гомоскедастичные, т. е. дисперсия ошибок постоянная, не зависит от значений предикторов;
3) отсутствует корреляция (автокорреляция) случайных ошибок разных наблюдений между собой: $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0,\ 1 \leq i < j \leq n$.

Итак, по набору из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, строим выборочное уравнение линейной регрессии $\hat{y}(x) = b_1 x + b_0$, где параметры b_0 и b_1 будут соответственно выборочными оценками параметров регрессии β_0 и β_1.

Подберем параметры b_1, b_0 так, чтобы прямая $\hat{y}(x) = b_1 x + b_0$ проходила как можно ближе к точкам (x_i, y_i), $i = 1, 2, \ldots, n$. Формализовать эту идею можно следующим образом. Рассмотрим функцию

$$F(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2.$$

Исследуем функцию F на минимум, приравняв частные производные к нулю:

$$\frac{\partial F}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_1 x_i - b_0) = 0,$$

$$\frac{\partial F}{\partial b_1} = 2\sum_{i=1}^{n}(y_i - b_1 x_i - b_0)x_i = 0.$$

Откуда получаем оценки:

$$b_0 = \frac{\sum\limits_{i=1}^{n} x_i^2 \sum\limits_{i=1}^{n} y_i \; - \; \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i y_i}{n\sum\limits_{i=1}^{n} x_i^2 \; - \; \Big(\sum\limits_{i=1}^{n} x_i\Big)^2},$$

$$b_1 = \frac{n\sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i \; - \; \sum\limits_{i=1}^{n} x_i y_i}{n\sum\limits_{i=1}^{n} x_i^2 \; - \; \Big(\sum\limits_{i=1}^{n} x_i\Big)^2}.$$

Из этих формул, в частности, следует, что для линейных моделей МНК-оценки являются линейными.

МНК-оценки для классической линейной регрессии являются несмещенными, состоятельными и наиболее эффективными оценками в классе всех линейных несмещенных оценок[14].

Теперь поясним, что означают вышеперечисленные требования. Разберем проблемы, возникающие в реальных задачах, когда эти требования не выполняются, и обсудим способы решения этих проблем.

Независимость факторов и случайных ошибок называется условием экзогенности. Если это условие не выполняется, то можно считать, что практически любые оценки будут крайне неудовлетворительными: они не будут даже состоятельными (т. е. точность оценок не улучшается при увеличении объема выборки).

При проведении регрессионного анализа исследователь может столкнуться с гетероскедастичностью случайных остатков. Гетероскедастичность (англ. *heteroscedasticity*) — это свойство случайных остатков, означающее неоднородность наблюдений, выражающуюся

[14]Требования к модели и свойства оценок устанавливает теорема Гаусса–Маркова. В англоязычной литературе иногда употребляют аббревиатуру BLUE (Best Linear Unbiased Estimator) — наилучшая линейная несмещенная оценка.

в неодинаковой (непостоянной) дисперсии случайной ошибки регрессионной модели. Гетероскедастичность противоположна гомоскедастичности, означающей однородность наблюдений, т. е. постоянство дисперсии случайных ошибок модели.

МНК-оценки параметров моделей являются несмещенными и состоятельными даже при гетероскедастичности, значит, при достаточном количестве наблюдений возможно применение обычного МНК. Однако наличие гетероскедастичности случайных ошибок приводит к неэффективности оценок, полученных с помощью метода наименьших квадратов. Кроме того, в этом случае оказывается смещенной и несостоятельной полученная по методу наименьших квадратов классическая оценка ковариационной матрицы оценок параметров. Следовательно, статистические выводы о качестве таких оценок могут быть неадекватными. В связи с этим тестирование на гетероскедастичность является одной из необходимых процедур при построении регрессионных моделей.

Наличие гетероскедастичности можно заметить на графиках остатков регрессии по некоторым переменным, по оцененной зависимой переменной или по номеру наблюдения. На этих графиках разброс точек может меняться в зависимости от значения переменных.

Для более строгой проверки применяют, например, статистические тесты Уайта, Голдфелда–Куандта, Бройша–Пагана, Парка, Глейзера, Спирмена.

Рассмотрим два метода снижения гетероскедастичности. Первый — взвешенный метод наименьших квадратов. Суть метода состоит в том, что каждое наблюдение взвешивается обратно пропорционально предполагаемому стандартному отклонению случайной ошибки в этом наблюдении. Такой подход позволяет сделать случайные ошибки модели гомоскедастичными. В частности, если предполагается, что стандартное отклонение ошибок пропорционально некоторой переменной Z, то ошибки делятся на эту переменную.

Второй метод — определение «областей компетенции» моделей, внутри которых дисперсия ошибки сравнительно стабильна, и использование комбинации моделей. Таким образом, каждая модель работает только в области своей компетенции, и дисперсия ошибки не превышает заданное граничное значение. Этот подход распространен в области распознавания образов, где часто используются сложные нелинейные модели и эвристики.

5.3. Пример построения парной линейной регрессии

Пусть даны наблюдения за двумя случайными величинами X и Y, которые мы будем трактовать как цену на нефть в долларах за баррель и доход бюджета в миллионах рублей. Требуется проверить, существует ли статистическая зависимость между этими величинами, и если да, то установить вид (уравнение) этой связи.

X	30	69	86	56	44	97
Y	204	267	313	207	130	320

X	53	66	39	29	34	31	92
Y	237	260	172	117	196	169	262

Данные внесем в таблицу (рис. 8). Переменные X и Y отнесем к порядковой шкале.

Исследование стоит начинать с графического изображения данных на плоскости XOY (см. рис. 2). Видно, что точки образовали некоторое облако и выстроились вдоль воображаемой наклонной прямой, т. е. подчинены линейной статистической связи.

Найдем уравнение этой прямой, используя для этого средства пакета SPSS.

Для построения уравнения линейной регрессии в пакете SPSS выбираем последовательно пункты меню `Analyze - Regression - Linear` (рис. 9). В ячейку ввода `Dependent` заносим переменную Y, а

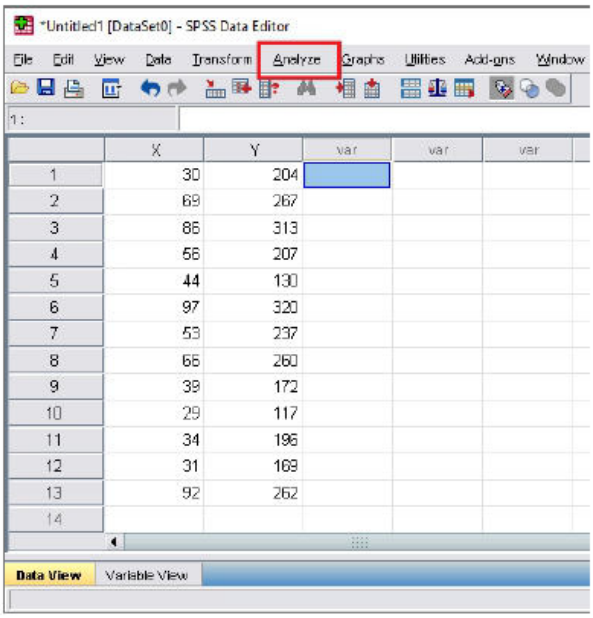


Рис. 8. Данные для построения уравнения парной линейной регрессии

в ячейку ввода `Independent(s)` — переменную X. В качестве метода ввода зависимой переменной в модель выбираем `Forward` (рис. 10); подробнее о методах ввода зависимой переменной в модель будет рассказано далее.

Обсудим полученные результаты, которые представлены в таблицах на рис. 11.

При проведении регрессионного анализа, во-первых, тестируется нулевая гипотеза об отсутствии связи между случайными величинами X и Y. В таблице ANOVA (рис. 11) представлены результаты промежуточных расчетов: различные суммы квадратов, которые используются для вычисления общей, остаточной и объясненной дисперсий, число степеней свободы и т. д. Все эти промежуточные расчеты не представляют большого интереса с практической точки зрения. Единственное важное число в таблице ANOVA — это величина p-value в столбце Sig., вероятность ошибки первого рода. В данном

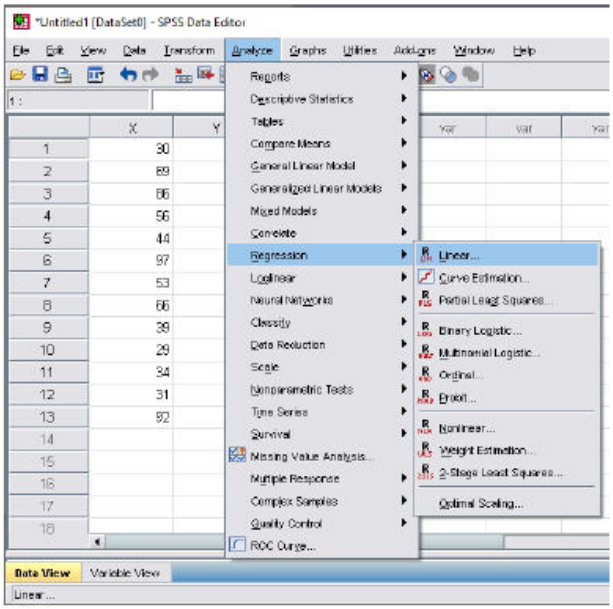


Рис. 9. Выбор пункта меню для построения парной линейной регрессии

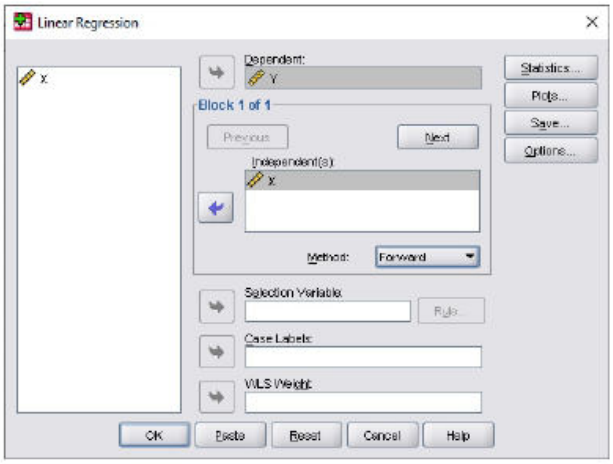


Рис. 10. Метод ввода зависимой переменной в модель

примере Sig. < 0.001, следовательно, на уровне значимости 0.001 нулевую гипотезу об отсутствии связи между случайными величинами X и Y отвергаем в пользу альтернативной, т. е. считаем связь между

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,870[a]	,758	,736	32,902

a. Predictors: (Constant), X

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37215,432	1	37215,432	34,378	,000[a]
	Residual	11907,798	11	1082,527		
	Total	49123,231	12			

a. Predictors: (Constant), X

b. Dependent Variable: Y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	91,696	23,636		3,879	,003
	X	2,289	,390	,870	5,863	,000

a. Dependent Variable: Y

Рис. 11. Результат построения линейной парной регрессии

X и Y статистически значимой на уровне значимости 0.001.

Если расчетная величина Sig. оказалась больше уровня значимости, например, Sig. $= 0.785$, нет оснований отвергнуть нулевую гипотезу, статистическая связь между переменными не обнаружена, и анализ заканчивают.

Далее если связь между переменными статистически доказана, то находят явный вид уравнения регрессии, беря коэффициенты из таблицы Coefficients. В этой таблице самые важные значения находятся в столбцах B и Sig. В столбце Sig. приведены результаты тестирования двух нулевых гипотез: о том, что константа и коэффициент при X в уравнении незначимо отличаются от нуля. Для константы Sig.$=$ $= 0.003$, для коэффициента при X Sig. < 0.001. Если, например, уровень значимости $\alpha = 0.05$, то обе нулевые гипотезы отвергаем как противоречащие экспериментальным наблюдениям. Значения константы и коэффициента при X смотрим в столбце B. Таким образом, уравнение регрессии имеет вид: $\hat{y}(x) = 2.289x + 91.696$.

Поскольку точные значения константы и коэффициента при x

неизвестны (они были оценены статистическими методами по выборочным данным), возникает вопрос о точности этих оценок. В столбце Std. Error приведены стандартные ошибки, которые равны 23.636 и 0.390 соответственно. Чем менее точки на графике разбросаны относительно прямой и чем больше наблюдений, тем ошибки меньше, соответственно оценки точнее.

После того как модель построена, необходимо ответить на вопрос, насколько эта регрессионная модель точна. В данном примере среднее значение Y равно 219.54, стандартное отклонение — 63.98, упрощенно говоря, можно сказать, что ожидаемое значение Y равно 219.54 ± 63.98. То есть не имея никакой априорной информации, можно сделать предположение о значении Y, при этом «коридор» ошибок составляет 63.98. Если же знать значение предиктора X, то можно сузить этот «коридор» на 75.8 %.

Более строго величина $R^2 = 0.758$, называемая коэффициентом детерминации, характеризует долю объясненной дисперсии (см. таблицу Model Summary). Коэффициент детерминации меняется от 0 до 1, чем он больше, тем точнее модель.

5.4. Множественная линейная регрессия

Очевидно, показатель, который необходимо спрогнозировать, может зависеть не от одного, а от многих факторов.

П р и м е р. Изучается зависимость уровня артериального давления (АД), а в качестве возможных факторов, оказывающих влияние, выбраны:

1) возраст;
2) индекс массы тела;
3) индекс курения;
4) количество минут, затрачиваемых на занятия спортом в день;
5) уровень холестерина.

Известно, что с возрастном АД, как правило, возрастает, однако пациент, ведущий здоровый активный образ жизни и следящий за

своим весом, может и в 70 лет иметь нормальное АД. В то же время курящий пациент с избыточным весом, ведущий малоподвижный образ жизни, более подвержен риску развития у него артериальной гипертензии уже в 40 лет.

Каждый фактор сам по себе лишь в незначительной степени может спрогнозировать уровень АД. Включение в модель нескольких факторов позволяет более полно охарактеризовать пациента и, следовательно, дать более точный прогноз касательно его АД.

Возникает резонный вопрос. Стоит ли вводить все имеющиеся у исследователя факторы (независимые переменные) в модель, чтобы объяснить наблюдаемые значения зависимой величины? Например, должен ли уровень холестерина входить в модель как фактор, влияющий на уровень АД? В общем случае ответ отрицательный — необоснованный ввод переменных в модель может ухудшить ее свойства.

Существуют специальные методы отбора факторов, которые надо вводить в модель. Суть их основана на оценке изменения скорректированного (англ. *adjusted*) коэффициента детерминации от ввода дополнительного фактора в модель. Эти методы встроены в стандартные статистические пакеты, например, SPSS.

Главная проблема применения (нескорректированного!) коэффициента детерминации R^2 состоит в том, что его значение увеличивается (точнее, не уменьшается) от введения в модель дополнительных переменных, даже если эти переменные никакого отношения к объясняемой переменной не имеют. В связи с этим сравнение моделей с неодинаковым числом факторов с помощью коэффициента детерминации, вообще говоря, некорректно. Для этих целей используют альтернативные показатели: скорректированный R^2 или различные информационные критерии (например, информационный критерий Акаике).

При введении дополнительного фактора в модель доля объясненной дисперсии увеличивается, однако штрафные множители («пла-

та» за количество факторов) тоже увеличиваются. В результате скорректированный R^2 увеличится лишь в том случае, если вновь вводимый фактор приводит к значительному росту доли объясненной дисперсии, т. е. может «объяснить» то, что не могли объяснить другие факторы.

Суть метода пошагового отбора в следующем:

1. Рассчитывается матрица корреляций и выбирается фактор, имеющий наибольшую корреляцию с зависимой переменной.
2. К выбранному регрессору последовательно добавляются каждый из оставшихся регрессоров и вычисляются скорректированные коэффициенты детерминации для каждой из моделей. К модели присоединяется тот регрессор, который обеспечивает наибольшее значение скорректированного R^2.
3. Процесс присоединения регрессоров прекращается, когда значение скорректированного R^2 становится меньше достигнутого на предыдущем шаге.

П р и м е р. Возраст, индекс массы тела, индекс курения и количество минут, затрачиваемых на физическую активность в неделю, значимо влияют на уровень АД человека. Пошаговое добавление этим факторов в модель позволит предсказать уровень АД пациента. С другой стороны, связь уровня холестерина или АЛТ с АД может быть столь незначительной, что присутствие этих переменных в модели лишь ухудшит ее свойства.

Введем понятие мультиколлинеарности. Мультиколлинеарность — наличие линейной зависимости между объясняющими переменными регрессионной модели. При этом различают полную коллинеарность, которая означает наличие функциональной линейной зависимости, и частичную или просто мультиколлинеарность — наличие сильной корреляции между факторами.

Полная коллинеарность приводит к неопределенности параметров в линейной регрессионной модели независимо от методов оценки.

Рассмотрим это на примере следующей линейной модели:

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \varepsilon.$$

Пусть факторы этой модели тождественно связаны следующим образом: $x_1 = x_2 + x_3$. Тогда рассмотрим исходную линейную модель, в которой к первому коэффициенту добавим произвольное число a, а из двух других коэффициентов это же число вычтем. Тогда имеем (без случайной ошибки): $y = (b_1 + a)x_1 + (b_2 - a)x_2 + (b_3 - a)x_3 = = b_1x_1 + b_2x_2 + b_3x_3 + a(x_1 - x_2 - x_3) = b_1x_1 + b_2x_2 + b_3x_3$.

Таким образом, несмотря на произвольное изменение коэффициентов модели, мы получили ту же модель. Такая модель принципиально неидентифицируема. Неопределенность существует уже в самой модели. Если рассмотреть трехмерное пространство коэффициентов, то в этом пространстве вектор истинных коэффициентов в данном случае не единственный, а представляет собой плоскость. Любая точка этой плоскости — истинный вектор коэффициентов.

Проблема полной коллинеарности факторов встречается редко. На практике чаще возникает другая ситуация — сильная корреляция между факторами.

Рассмотрим пример титрования дозы препарата у детей до 7 лет в зависимости от возраста и веса ребенка. Чем ребенок старше и чем он больше весит, тем большую дозу ему необходимо назначить. Желая учесть оба фактора, исследователь включил (не используя методы пошагового отбора переменных!) их в модель и с помощью регрессионного анализа построил следующую модель:

$$\text{доза} = 2 \times \text{вес} + 1 \times \text{возраст}.$$

Пусть все величины, входящие в модель, обезразмерены. Коэффициенты при независимых величинах имеют содержательный смысл: при увеличении веса на 1 кг надо увеличивать дозу на 2 у. е., и с каждым следующим годом надо увеличивать дозу на 1 у. е. Из-

вестно, что вес ребенка сильно коррелирует с его возрастом, а потому включение в модель и возраста, и веса как предикторов может (хотя это не обязательно) привести к неверной идентификации параметров. Действительно, если вес $\approx$ возраст, то при незначительных изменениях в исходных данных программа могла построить модель вида:

$$\text{доза} = 4 \times \text{вес} - 1 \times \text{возраст}.$$

Полученная модель противоречит здравому смыслу. Именно содержательный смысл отрицательного коэффициента при переменной «возраст» состоит в том, что доза препарата тем меньше, чем старше ребенок. Это и есть проявление мультиколлинеарности, следствием которой является плохая идентифицируемость параметров.

П р и м е р. Интраоперационная кровопотеря является риском развития острого делирия в ближайшем послеоперационном периоде. Желая спрогнозировать риски и повысить точность оценок, исследователь строит модель и вводит в нее уровень гемоглобина, уровень гематокрита и число эритроцитов. Проблема такой модели — мультиколлинеарность, так как все эти три фактора очень сильно коррелированы между собой. В результате оценки параметров окажутся неточными (или вовсе абсурдными), а качество прогноза сильно ухудшится.

В модель должны попадать только значимые независимые предикторы. Для этого, например, в программе SPSS нужно в соответствующем диалоговом окне выбрать метод `Forward LR` или `Stepwise`.

5.5. Нелинейная связь величин

Для точного описания уравнения регрессии (необязательно линейной) требуется знать условный закон распределения случайной величины Y, чтобы найти ее условное математическое ожидание $M[Y|X = x]$. В статистической практике такую информацию, как

правило, не удается получить, а потому на основе наблюдаемых данных выбирают подходящую аппроксимацию функции $M[Y|X=x]$.

Поясним, какая связь между следующими тремя сущностями:

1) истинная (вообще говоря, нелинейная) регрессия $f(x) = M[Y|X=x]$. Она была бы построена, если бы была известна плотность совместного распределения;
2) линейная аппроксимация функции регрессии $\tilde{f}(x) \approx f(x)$. Эта функция тоже могла быть построена, если бы была известна плотность совместного распределения;
3) оценка линейной аппроксимации функции регрессии $\hat{f}(x)$.

Пусть случайная величина Y связана с X равенством $Y = X^{1.2} + \varepsilon$, где X равномерно распределена на $[1,3]$; ε — нормально распределенная случайная величина с нулевым математическим ожиданием и независящей от X дисперсией.

Истинная регрессия имеет вид: $y = f(x) = M[Y|X=x] = x^{1.2}$.

График функции $y = f(x) = x^{1.2}$, $x \in [1,3]$, близок к линейному. Учитывая, что линейные модели предпочтительны в смысле своей простоты, разумным представляется построить аппроксимацию функции $f(x)$ в классе линейных:

$$\tilde{f}(x) = \beta_1 x + \beta_0.$$

Параметры выбираются исходя из приближения в той или иной норме, например,

$$\max_{x\in[1,3]} \left| x^{1.2} - \beta_1 x - \beta_0 \right| \xrightarrow[\beta_0,\,\beta_1]{} \min,$$

$$\int_1^3 (x^{1.2} - \beta_1 x - \beta_0)^2\, dx \xrightarrow[\beta_0,\,\beta_1]{} \min.$$

Ошибка (систематическая), возникающая при этом, будет небольшой, а работать с функцией $\tilde{f}(x)$ может быть проще, чем с $f(x)$.

В реальных задачах точный вид взаимосвязи случайных величин нам вообще неизвестен, имеется лишь конечный набор наблюдений

за парой X, Y. Расположение точек дает основание предположить линейную взаимосвязь X и Y и построить выборочное уравнение $\hat{f}(x) = b_1 x + b_0$, которое по вероятности будет сходиться к $\tilde{f}(x) = \beta_1 x + \beta_0$ при неограниченном увеличении объема выборки n.

Важно понимать, что поскольку мы ошиблись в выборе класса функций, когда строили выборочное уравнение (что на практике, увы, нередко происходит), получаемые оценки не будут состоятельными. То есть как бы мы ни увеличивали объем выборки, выборочная оценка $\hat{y}$ не будет сходиться к истинной функции регрессии $f(x)$.

Задания для самостоятельной работы

1. Профессор Р. А. Шамойлова учит студентов проводить регрессионный анализ. На лекции она сказала: «Основной предпосылкой регрессионного анализа является то, что только результативный признак подчиняется нормальному распределению, а факторные признаки могут иметь произвольный закон распределения». Согласны ли вы с этим утверждением?

2. Кроме метода наименьших квадратов существуют и другие методы оценивания. Пусть X, Y подчиняются двумерному нормальному распределению. Проведите оценку параметров уравнения линейной регрессии $y = \beta_1 x + \beta_0$ по методу максимального правдоподобия. Отличаются ли результаты от тех, что были получены методом наименьших квадратов?

3. При построении линейной регрессии в п 5.2 по набору из n наблюдений (x_i, y_i), $i = 1, 2, \ldots, n$, за X и Y мы предполагали, что не все x_i равны между собой. Зачем это требование нужно? Как геометрически интерпретировать проблемы, которые возникнут в случае его нарушения?

4. Что такое мультиколлинеарность? К каким негативным последствиям она приводит? Какие существуют методы борьбы с мультиколлинеарностью?

5. В чем суть информационных критериев Акаике и Байеса? Для чего они применяются? В чем их отличие?

6. Логистическая регрессия

Линейная регрессия не всегда способна качественно описать и предсказать значения зависимой переменной. Выбирая для построения модели линейное уравнение, мы естественным образом не накладываем никаких ограничений на значения зависимой переменной, однако эти ограничения могут быть существенными. Как следствие, линейная регрессионная модель может дать бессмысленные результаты.

П р и м е р. Проводится трансплантация почки, требуется оценить вероятность гибели почки в течение одного года после операции (например, в результате отторжения, ишемии или гибели пациента). Данная операция имеет лишь два возможных исхода: почка функционирует или погибла.

В качестве предикторов исхода трансплантации используется ряд факторов: возраст донора и реципиента, время тепловой и холодовой ишемии почки, длительность пребывания донора в реанимации и т. д. Если строить модель в классе линейных, т. е. вида $p = a_1x_1 + a_2x_2 + \ldots + a_mx_m + a_0$, где p — это вероятность гибели почки в течение года после операции; $x_i, i = 1, 2, \ldots, m$, — это предикторы исхода, $a_i, i = 0, 1, 2, \ldots, m$, — соответствующие коэффициенты, то при определенных значениях предикторов вероятность, вычисляемая подобным образом, будет либо отрицательной, либо больше единицы, что невозможно по определению.

С целью решения данных проблем рассматривают иной класс уравнения регрессии. Пусть p — это вероятность прогнозируемого события; $x_i, i = 1, 2, \ldots, m$, — предикторы исхода; $a_i, i = 0, 2, \ldots, m$, — соответствующие коэффициенты.

Введем в рассмотрение ненаблюдаемую величину

$$y = a_1x_1 + a_2x_2 + \ldots + a_mx_m + a_0,$$

которую будем трактовать как «направленность» к событию: чем

больше y, тем больше вероятность события. При этом мы не накладываем ограничений на y, который может принимать значения вне интервала $[0; 1]$. Вероятность p будем оценивать следующим образом:

$$p = \frac{e^y}{1 + e^y}.$$

Справедливы следующие соотношения: $p \to 1$ при $y \to +\infty$ и $p \to 0$ при $y \to -\infty$.

Вообще, логистическая регрессионная модель предназначена для решения задач предсказания значения непрерывной зависимой переменной, при условии, что эта зависимая переменная может принимать значения на интервале от 0 до 1. В силу такой специфики ее часто используют для предсказания вероятности наступления некоторого события в зависимости от значений некоторого числа предикторов.

Можно использовать логистическую регрессию и для решения задач с бинарным откликом. Подобные задачи появляются, когда зависимая переменная может принимать только два значения (пациент выжил или не выжил, заемщик отдал долг или не отдал). По сути, речь идет о задаче классификации объектов на две группы.

Устанавливается пороговое значение, начиная с которого, считается, что событие скорее наступит. Далее вычисляется вероятность наступления искомого события по набору предикторов. Если расчетная вероятность превысила порог, объект относят к первой группе, иначе — ко второй.

Задание для самостоятельной работы

В книге А.Н. Толстого «Золотой ключик, или Приключения Буратино» есть эпизод:

«Сова приложила ухо к груди Буратино.

– Пациент скорее мертв, чем жив, — прошептала она и отвернула голову назад на сто восемьдесят градусов.

Жаба долго мяла влажной лапой Буратино. Раздумывая, глядела выпученными глазами в разные стороны. Прошлепала большим ртом:

— Пациент скорее жив, чем мертв...»

Проанализируйте решение Совы и Жабы с точки зрения логистической регрессии.

7. Анализ выживаемости

7.1. Введение в проблематику и основные понятия

Введение в проблематику теории выживаемости начнем с рассмотрения примера.

П р и м е р. Проводится трансплантация почки, исследуется выживаемость пациентов после операции, в частности, важно узнать, как проводимая иммуносупрессивная и противовирусная терапия влияют на выживаемость.

Очевидно, наиболее важной переменной является продолжительность жизни пациентов с момента трансплантации. В принципе для описания среднего времени жизни и сравнения нового метода иммуносупрессии со старыми можно было бы использовать стандартные параметрические и непараметрические методы (t-тест, тест Манна–Уитни). Однако анализируемые данные имеют существенную особенность, связанную с тем, как строится выборка.

Во-первых, пациентов включают в группу наблюдения на всем протяжении исследования, а живут многие пациенты после трансплантации десятилетиями, поэтому когда исследование заканчивается, многие пациенты еще живы. Соответственно, истинная продолжительность жизни этих пациентов остается неизвестной. Естественно, не хотелось бы терять ту информацию, которую удалось о них собрать. Если пациент после трансплантации наблюдается в течение 10 лет (и жив на момент окончания исследования), то это сама по себе ценная информация, свидетельствующая в пользу данного метода лечения.

Во-вторых, исследователь может потерять пациента из виду, если тот переехал в другой город или сменил клинику. Пациент также может умереть с функционирующим трансплантатом в результате несчастного случая. Во всех этих примерах время, в течение которого пациент способен жить после трансплантации, остается неизвестным. Таких пациентов называют выбывшими.

Наблюдения, которые содержат неполную информацию о времени жизни, называются цензурированными наблюдениями. Обсудим механизмы цензурирования.

При фиксированном цензурировании выборка из n объектов наблюдается в течение фиксированного времени. Число объектов, для которых наступает терминальное событие, или число смертей, случайно, но общая продолжительность исследования фиксирована. Так, в нашем примере исследование проводилось в течение 20 лет — это и есть фиксированный промежуток времени. Каждый объект имеет максимально возможный период наблюдения $\tau_i, i = 1, \ldots, n$, который может варьироваться у разных объектов (поскольку некоторые пациенты включаются в исследование позже других), однако фиксирован заранее (не раньше, чем с момента начала наблюдения, и до конца исследования). Вероятность того, что объект i будет жив в конце своего периода наблюдения, равна $S(i)$, а общее число смертей является случайным.

При случайном цензурировании выборка из n объектов наблюдается столь долго, сколько это необходимо, чтобы d, $d \leq n$, объектов испытали событие. В этой схеме число смертей d, которое определяет точность исследования, фиксировано заранее, и его можно использовать в качестве параметра. Недостатком данного подхода является то, что в этом случае общая продолжительность исследования случайна и не может быть точно известна заранее. Примером такого подхода мог бы быть проспективный эксперимент с лабораторными животными, которых единовременно подвергают воздействию и далее наблюдают за их выживаемостью до тех пор, пока половина животных не погибнет. Таким образом исследователь может оценить срединное время жизни.

При проведении цензурирования можно также указать направление, в котором производится цензурирование. Цензурирование справа имеет место, если известно, в какой момент эксперимент был начат и что он закончится в момент времени, расположенный справа

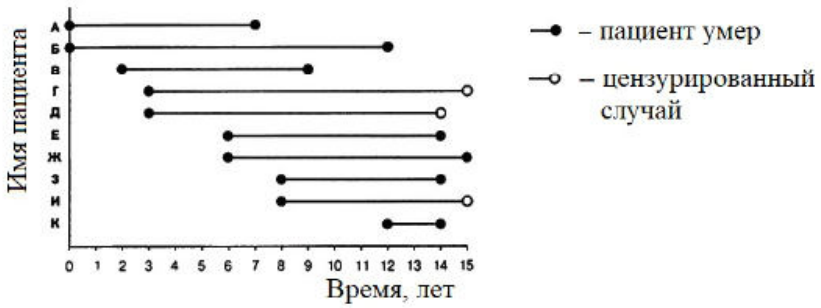


Рис. 12. График выживаемости пациентов

от точки начала эксперимента. Например, проведена трансплантация почки (дата трансплантации известна), пациента наблюдали в посттрансплантационном периоде пять лет, и исследование закончилось. Можно сказать, что пациент прожил не менее пяти лет с момента трансплантации. Имеет место цензурирование справа.

Рассмотрим вариант, когда нет информации о времени начала эксперимента. Например, у исследователя есть данные о том, когда пациент поступил в клинику и что он выздоровел через три недели после госпитализации. Однако при этом может отсутствовать информация, когда впервые проявились симптомы его заболевания. Таким образом, можно лишь утверждать, что заболевание длилось не менее трех недель. Здесь имеет место левое цензурирование. Цензурирование также может быть двусторонним.

Цензурирование можно классифицировать как однократное или многократное. Однократное цензурирование происходит в один момент времени (эксперимент заканчивается спустя некоторое фиксированное время, см. выше пример с лабораторными животными). С другой стороны, в биомедицинских исследованиях естественным образом возникает многократное цензурирование, например, когда пациенты выписываются из стационара, пройдя курс лечения в различных объемах (или разной продолжительности), и исследователю лишь известно, что пациент дожил до соответствующего момента цензурирования.

На рис. 12 показан ход исследования. Наблюдение за пациентом представлено горизонтальным отрезком. Левый конец отрезка — это начало наблюдения. На правом конце отрезка находится черный или белый кружок. Черный кружок означает, что пациент умер (произошел исход), и, таким образом, продолжительность его жизни известна. Белый кружок означает, что исследование закончилось до смерти пациента либо он выбыл из наблюдения. Относительно выбывших известно только, что они прожили не меньше определенного срока.

Методы анализа выживаемости в основном применяются к тем же статистическим задачам, что и другие методы, однако их особенность состоит в том, что они используются при наличии цензурированных или, как иногда говорят, неполных данных. Отметим также, что вместо традиционной функции распределения в этих методах, как правило, используется так называемая функция выживания, представляющая собой вероятность того, что объект проживет время больше t. Построение таблиц времен жизни, подгонка распределения выживаемости, оценивание функции выживания с помощью процедуры Каплана–Мейера являются описательными методами исследования цензурированных данных. Некоторые из названных методов позволяют также сравнивать выживаемость в двух и более группах. Кроме того, используя аппарат анализа выживаемости, можно строить регрессионные модели для оценивания зависимостей между многомерными непрерывными переменными и временем жизни.

7.2. Анализ таблиц времен жизни

Наиболее естественным способом описания выживаемости в выборке является построение таблиц времен жизни. Это один из старейших методов анализа данных о выживаемости (времен отказов) (см., например, [7–9]). Такую таблицу можно рассматривать как «расширенную» таблицу частот. Область возможных времен наступления критических событий (смертей, отказов и др.) разбивается на K интервалов: $[t_0, t_1], [t_1, t_2], \ldots, [t_{K-1}, t_K]$. Для каждого интервала

вычисляется число и доля объектов, которые в начале рассматриваемого интервала были живы, число и доля объектов, которые умерли в данном интервале, а также число и доля объектов, которые были изъяты или цензурированы в каждом интервале. На основании этих величин вычисляются некоторые дополнительные статистики.

Число объектов, изучаемых в течение i-го интервала, — r_i. Это число объектов, которые были живы в начале рассматриваемого временного интервала, минус половина числа изъятых или цензурированных объектов.

Доля умерших в течение i-го интервала — q_i. Это отношение числа объектов, умерших в соответствующем интервале, к числу объектов, изучаемых на этом интервале.

Доля выживших в течение i-го интервала — $p_i = 1 - q_i$. Эта доля равна единице минус доля умерших.

Кумулятивная доля выживших к концу i-го интервала — P_i (функция выживания). Это кумулятивная доля выживших к концу соответствующего временного интервала. Поскольку вероятности выживания считаются независимыми на разных интервалах, эта доля равна произведению долей выживших объектов на текущем и всех предыдущих интервалах: $P_i = P_{i-1}\, p_i = p_i\, p_{i-1} \ldots p_1$; считается, что $P_0 = 1$. Полученная доля как функция от времени называется также выживаемостью или функцией выживания (точнее, это оценка функции выживания).

Плотность вероятности гибели — это оценка вероятности гибели в соответствующем интервале, определяемая следующим образом: $f_i = (P_{i-1} - P_i)/h_i$, где f_i — оценка вероятности гибели в i-м интервале; $h_i = t_i - t_{i-1}$ — ширина i-го интервала.

Функция интенсивности определяется как вероятность того, что объект, выживший к началу соответствующего интервала, умрет в течение этого интервала. Оценка функции интенсивности вычисляется как число отказов, приходящихся на единицу времени соответствующего интервала, деленное на среднее число объектов, дожив-

ших до момента времени, находящегося в середине интервала.

Медиана ожидаемого времени жизни — это точка на временной оси, в которой кумулятивная функция выживания равна 0.5. Другие процентили (например, 25-й и 75-й процентиль, или квартили) кумулятивной функции выживания вычисляются по такому же принципу. Отметим, что 50-й процентиль (медиана) кумулятивной функции выживаемости обычно не совпадает с точкой выживания 50 % выборочных наблюдений. Совпадение происходит тогда и только тогда, когда за прошедшее к этому моменту время не было цензурированных наблюдений.

Для получения надежных оценок трех основных функций (функции выживания, плотности вероятности и функции интенсивности) и их стандартных ошибок на каждом временном интервале необходимо, чтобы объем выборки был достаточный.

Задания для самостоятельной работы

1. Проводится анализ выживаемости пациентов после резекции печени в пределах сегмента. Выборка состоит из 20 пациентов. У 10 пациентов наступил летальный исход, еще 10 случаев были цензурированы. Повлияет ли на оценку доли выживших пациентов порядок, в котором наступали цензурированные и летальные исходы? Если да, то каким образом?

2. Василий возглавляет отдел анализа данных в крупном банке. Перед отделом стоит задача изучения оттока клиентов. В таблице данных имеются записи как о клиентах, которые отказались от услуг банка, так и о клиентах, которые на данный момент продолжат обслуживание в банке.

Василий не изучал прикладную статистику и предлагает, проводя анализ времени обслуживания, удалить все записи о клиентах, которые не отказались от услуг банка. Василий аргументирует свое решение тем, что для данных клиентов изучаемое событие не наступило, а потому никакую полезную информацию из соответствующих записей извлечь нельзя. Согласны ли вы с Василием? Будут ли оценки. полученные по методу Василия, несмещенными?

8. Дизайны исследования

Все статистические исследования можно разделить на две группы в зависимости от решаемых задач: описательные и анализирующие причинно-следственные связи. Последние в медицине называются клиническими.

8.1. Описательные исследования

В описательных исследованиях можно выделить такие исследования, как описание случая и описание серии случаев. Единственным отличием данных исследований является количество пациентов. Если в исследование входит один пациент, т. е. один случай какого-либо заболевания или состояния, в такой ситуации проводится описание случая. Если в исследование входит несколько пациентов, т. е. несколько случаев заболевания или состояния, в такой ситуации проводится описание серии случаев. Два данных дизайна разделены в связи с тем, что их реализация осуществляется различными способами с математической точки зрения. При описании случая в публикации приводятся конкретные значения различных параметров пациента. При описании серии случаев, как правило, применяются статистические параметры, агрегирующие данные обо всех входящих в исследование пациентах (средние значения, медианы, проценты и др.) и характеризующие разброс показателей пациентов (стандартные отклонения, квартили, ошибки процентов и др.).

Необходимо отметить, что описание случая и описание серии случаев, как правило, имеют низкий уровень доказательности. Такие исследования проводятся обычно при обнаружении редкого случая или нескольких редких случаев какого-либо заболевания или состояния.

8.2. Клинические исследования

Клинические исследования, основной целью которых является не просто описание одного или нескольких случаев заболевания или состояния, а выяснение различных причинно-следственных связей (между факторами риска и возникновением заболевания, применяемыми препаратами и излечением, методами диагностики и диагностическими ошибками и т. д.), в свою очередь, также разделяются на две подгруппы — экспериментальные и наблюдательные (обсервационные) исследования.

Основным отличием экспериментальных исследований от наблюдательных является то, что при их реализации на исследуемые субъекты (пациентов, экспериментальных животных) оказывается какое-либо воздействие, а при реализации наблюдательных исследований лишь фиксируется развитие ситуации, которая никак не зависит от исследователя и его вмешательство возможно лишь в критических ситуациях, когда пациентам грозит опасность.

Экспериментальные клинические исследования

Экспериментальные исследования обычно проводятся при изучении различных методов лечения (препараты, схемы лечения, хирургические методы и т. д.). Данные исследования, как правило, не являются диссертационными, так как их проведение, как правило, требует довольно существенных временных и финансовых затрат. При этом экспериментальные клинические исследования могут быть рандомизированными или нерандомизированными.

1. Нерандомизированные экспериментальные клинические исследования. В целом данный тип исследования довольно существенно уступает по уровню доказательности рандомизированным экспериментальным клиническим исследованиям, он может иметь место в тех случаях, когда, например, рандомизация невозможна.

Дизайн нерандомизированного экспериментального клинического исследования довольно прост. Набирается две максимально оди-

наковые (по половозрастной структуре, по степени заболевания и т. д.) группы пациентов. Первая группа получает один препарат, одну схему или хирургический метод лечения, а вторая группа – второй препарат, вторую схему или хирургический метод лечения. Полученные результаты лечения (число излеченных пациентов, нежелательных реакций, смертельных исходов, изменение лабораторных данных и т. д.) сравниваются. Выявленные различия могут быть интерпретированы как большая эффективность той или иной схемы лечения.

Низкий уровень доказательности данных исследований связан с тем, что как сами пациенты, так и исследователи могут, даже несознательно, смещать результаты исследования, приводя к систематическим или случайным ошибкам. Это может происходить на этапе формирования исследуемых групп, например, когда в одну группу включаются пациенты с более легкой степенью тяжести изначального состояния. Также смещение результатов исследования может происходить на этапе оценки результатов лечения, например, когда врач, зная, какая группа пациентов получает новый исследуемый препарат, может интерпретировать незначительные изменения как улучшение состояния пациентов.

Необходимо отметить, что подобные исследования требуют использования уже не только описательных статистических параметров, но и методов сравнительной статистики, что позволяет свидетельствовать о статистически значимых различиях между группами в процессе или в конце исследования. Применение подобных методов позволяет исследователям сделать заключение о причинности или случайности изучаемых медицинских закономерностей.

2. Рандомизированные экспериментальные клинические исследования. Рандомизированные экспериментальные клинические исследования призваны устранить одну из возможных причин смещения результатов лечения – несопоставимость исследуемых групп пациентов. Потенциальное смещение результатов исследования, связанное

с данной причиной, устраняется с помощью процесса рандомизации. Рандомизация является методом распределения пациентов на группы и заключается в том, что пациенты распределяются на группы максимально случайно. Чаще всего используются три основных метода рандомизации: таблицы случайных чисел, математические алгоритмы генераторов псевдослучайных чисел и физические способы рандомизации, такие как монетки, жеребьевка и др.

3. Рандомизированные плацебо-контролируемые экспериментальные клинические исследования. Рандомизированные плацебо-контролируемые экспериментальные клинические исследования являются золотым стандартом клинических исследований и имеют наивысший уровень доказательности. Подобные исследования помимо рандомизации включают еще два важных аспекта — участие контрольной группы и метод ослепления.

Помимо одной или нескольких исследуемых групп, которые получают различные экспериментальные схемы лечения, в исследование включается контрольная группа пациентов, которая получает либо плацебо, либо уже известную и широко применяемую схему лечения. Если в процессе исследования изучаются пациенты, состояние которых может угрожать их жизни и здоровью, то, естественно, плацебо в таком случае не применяется, а контрольная группа должна получать известную широко используемую схему лечения. Если же изучаются пациенты, состояние которых не является жизнеугрожающим, то предпочтительно применение плацебо.

Плацебо — муляж препарата или набора препаратов, по органолептическим свойствам схожих с препаратами, которые получают исследуемые группы пациентов, при этом данные вещества не оказывают никакого фармакологического и терапевтического эффекта.

Метод ослепления заключается в исключении предвзятого отношения к состоянию самих пациентов и исследователей. Ослепление может быть простым, двойным и тройным. При простом ослеплении пациенту не сообщается, какой из исследуемых препаратов или

плацебо он получает. При двойном ослеплении об этом не знают ни пациент, ни специалист, который непосредственно оценивает результат лечения, при тройном ослеплении об этом не знают ни пациент, ни исследователь, ни специалист, который непосредственно оценивает результаты лечения.

4. Рандомизированные неконтролируемые экспериментальные клинические исследования. Данный тип исследований отличается от предыдущего лишь тем, что в нем не участвует контрольная группа и соответственно не применяется метод ослепления. В связи с этим уровень доказательности данных исследований существенно ниже, чем у рандомизированных контролируемых экспериментальных клинических исследований. Поэтому к данному типу исследований по возможности лучше не прибегать.

Наблюдательные (обсервационные) клинические исследования

В зависимости от того, что берется за основу формирования исследуемых групп, некомбинированные наблюдательные (обсервационные) исследования подразделяются на исследования «случай–контроль», когортные и кросс-секционные.

1. Исследование «случай–контроль» — исследование, при котором изучаемые группы пациентов набираются по следующему принципу. В группу I входят пациенты, имеющие изучаемое заболевание или состояние (группа случаев), в группу II — пациенты, не имеющие изучаемого заболевания или состояния (группа контроля). После набора групп у пациентов выясняется наличие изучаемых факторов риска и производится оценка наличия данных факторов в группе случаев и группе контроля. Данное исследование является ретроспективным, так как факт наступления заболевания известен, влияние факторов риска зафиксировано уже на этапе начала исследования, а исследователь только собирает эту информацию.

Исследование «случай–контроль» имеет следующие достоинства:

1) число включенных в исследование лиц намного меньше, чем в когортном исследовании;
2) низкая стоимость в связи с тем, что нет необходимости долго наблюдать большое число исследуемых;
3) результаты могут быть известны практически сразу после получения данных, необходимо только время на их статистическую обработку;
4) можно оценить связь между заболеванием и практически неограниченным числом факторов риска.

Недостатки исследования по типу «случай—контроль»:

1) можно получить лишь приближенную оценку показателя относительного риска в виде показателя отношения шансов;
2) можно оценить воздействие факторов риска только на одно заболевание или состояние.

2. Когортные исследования — исследования, при которых изучаемые группы пациентов набираются по следующему принципу. В группу I входят пациенты, не имеющие изучаемого заболевания или состояния, но имеющие изучаемый фактор риска (экспонированная группа), в группу II – пациенты, также не имеющие изучаемого заболевания или состояния, но не имеющие еще и изучаемый фактор риска (неэкспонированная группа).

Как правило, когортные исследования ассоциируются с проспективным типом исследований, когда после набора групп пациенты наблюдаются какое-то время (зависит от изучаемого заболевания и может составить десятки лет) и в конце исследования производится оценка частоты развития изучаемого заболевания или состояния в экспонированной и неэкспонированной группах пациентов. Однако когортные исследования могут проводиться ретроспективно, когда исходы пациентов уже известны, но формирование групп тем не менее осуществляется по принципу наличия или отсутствия факторов.

Естественно, когортное исследование также имеет свои достоинства и недостатки. Основные достоинства когортных исследований:

1) когортное исследование — единственный способ истинной оценки относительного риска. Относительный риск — числовая характеристика, позволяющая оценить силу связи между выраженностью фактора риска с частотой развития заболевания — может быть рассчитан только при проведении когортного исследования. При проведении исследования «случай–контроль» данный показатель рассчитать нельзя, рассчитывается только отношение шансов. В случае редких заболеваний отношение шансов дает приблизительную оценку (хотя, как правило, довольно точную) относительного риска;
2) имеется возможность оценки влияния изучаемого фактора риска на различные заболевания или состояния. В процессе наблюдения могут развиться не только изучаемые заболевания или состояния, но и другие, возникновение которых не предполагалось в начале исследования.

Недостатки когортных исследований:

1) число включенных в исследование лиц должно быть значительно больше, чем число пациентов с изучаемым заболеванием (характерно для проспективных когортных исследований). Несомненно, не у всех, кто будет участвовать в исследовании, разовьется изучаемое заболевание, поэтому необходимо включать в исследование большее количество человек, для того чтобы количество пациентов с развившимся заболеванием также было приемлемым;
2) высокая стоимость исследования из-за того, что приходится исследовать большое число людей в течение продолжительного времени (характерно для проспективных когортных исследований): изучаемые заболевания могут развиваться годами и десятилетиями;

3) результаты долгое время остаются неизвестными (характерно для проспективных когортных исследований); они станут известны только в конце исследования;
4) позволяют оценить связь между заболеванием и воздействием относительно небольшого числа факторов (определенных в начале исследования).

3. Кросс-секционные (поперечные) исследования — проводятся одномоментно для выяснения распространенности факторов риска и исходов. Необходимо отметить, что данное исследование, как правило, не проводится для выяснения причинно-следственной связи между факторами риска, лечением, исходами и т. д.

4. Исследование временных серий — заключается в изучении одной группы пациентов, сформированной по какому-либо принципу, в динамике. Как правило, после формирования данной группы обозначаются контрольные временные точки (раз в неделю, месяц или год), когда пациенты вновь обследуются, и исследователь получает о них новый набор данных. В дальнейшем исследователем применяются специализированные статистические методы, позволяющие установить различия в динамике изучаемых показателей (парные методы или методы повторных измерений).

Задания для самостоятельной работы

1. Приведите пример, который показывает, что для дизайна «случай—контроль» невозможно корректно рассчитать относительный риск. Можно ли для этого дизайна корректно рассчитать отношение шансов? Почему?

2. Проводится клиническое исследование результатов эндоваскулярной коррекции стеноза почечной артерии. Сформировано две группы: в группе исследования пациентам проводится ангиопластика, в группе контроля пациенты получают только медикаментозную терапию. В группу исследования попали пациенты со стенозом более 70 %, а группу контроля — оставшиеся пациенты. Можно ли назвать данное исследование рандомизированным?

Список библиографических ссылок

1. *Гмурман В. Е.* Теория вероятностей и математическая статистика : учеб. пособие для вузов. 9–е изд., стер. М. : Высш. шк., 2003. 479 с.
2. *Rubin D.* Inference and Missing Data // Biometrika. 1976. № 3. P. 581–592.
3. *Cheema J.* A Review of Missing Data Handling Methods in Education Research // Review of Educational Research. 2014. № 4. P. 487–508.
4. *Peugh J., Enders C.* Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement // Review of Educational Research. 2004. № 4. P. 525–556.
5. *Enders C.* Applied Missing Data Analysis. N. Y., L., 2010. 377 p.
6. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М. : Наука, 1973. 899 с.
7. *Berkson J., Gage R.* Calculation of survival rates for cancer // Proc. Staff. Meet. Mayo. Clin. 1950. Vol. 25, № 11. P. 270–86.
8. *Cutler S. J., Ederer F.* Maximum utilization of the life table method in analyzing survival // J. Chronic Dis. 1958. Vol. 8, № 6. P. 699–712.
9. *Gehan E. A.* Estimating survival functions from the life table // J. Chronic Dis. 1969. Vol. 21, № 9. P. 629–644.