Тема:

«Количественные исследования. Использование инструментов статистики и прогнозирования при принятии решений»

MBA Start
Бизнес-образование без границ

Конспект видеолекции

Оглавление

Введение

Данный курс имеет своей целью общее знакомство слушателей с несколькими важными методами, применяющимися в бизнесе. К применению будут предложены сравнительно несложные формулы, принимаемые без доказательств. Студенты, которым требуются более точные, математически выверенные формулировки, могут найти все необходимое в соответствующей литературе.

Разумеется, курс не претендует на полноту: математические методы, используемые в бизнесе, развивались многие столетия и для самых разных — порой, весьма неожиданных — целей (достаточно вспомнить, что основы интегрального исчисления, например, зародились вследствие желания винодела знать объем винной бочки с достаточной точностью). Количество литературы, накопленной за это время, невозможно изучить в течение нескольких жизней — что уж говорить об одной. Заинтересованный слушатель может найти более подробную информацию в книгах; автор же считает своей главной задачей подсказать направления для поиска.

Раздел 1. Основы статистики и регрессионного анализа

1.1. Важность подготовки данных и знания закономерностей

Данные совершенно необходимы в любом бизнесе. Даже если вы не планируете пользоваться другими методами, вам необходимо знать, как меняется бизнес-среда, как меняется положение в ней вашей компании, как исполняются ваши планы и т.д. Если же вы планируете использовать другие методы, тогда подробная статистика необходима вам буквально как воздух. Тут следует отметить, что в условиях действующего бизнеса — а особенно, малого бизнеса - стоимость такой статистики может быть запредельно высока; поэтому, принимая решение о ведении статистики, имеет смысл предварительно оценить стоимость такого начинания и каким-то образом ограничить набор записываемых величин — возможно, полный набор данных вам и не понадобится. Например, в кафе вполне можно вести урезанную статистику: количество посетителей в зале, средний чек и время, проведенное посетителем в кафе, — не учитывая при этом точный набор и порядок блюд, количество посетителей за одним столом и т.д. (разумеется, если ваше кафе имеет специфическое позиционирование, то набор отслеживаемых параметров может быть и другим).

Однако зачастую, накопив огромное количество данных, компания просто не знает, что с ними делать. Вы, возможно, чувствуете, что в понедельник пиво в вашем кафе пользуется повышенным спросом, — но ни проверить, ни опровергнуть это утверждение вы не можете: ведь сотрудники каждый раз говорят вам, что это все «случайно» (разумеется, случайно — это вы и сами знаете, но хотели бы понять закономерности, лежащие в основе таких случайностей). При этом речь о более сложных выводах не идет; для начала, необходимо разобраться с самыми простыми вещами.

Занимаясь **описательной статистикой**, мы должны понимать, что это «первая лига», в то время как инференционная статистика (статистика выводов) — соответственно, «высшая». Подготовка данных при этом, пожалуй, — «вторая лига». Но, если продолжать футбольно-хоккейные ассоциации, именно так закладывается основа для удачной игры «сборной».

В качестве «затравки» можно привести следующий пример: представьте себе, что вы продавец газет (точнее, одной газеты). Вы покупаете газету за 1 доллар, продаете за 2. При этом редакция газеты готова забрать все непроданные экземпляры по 50 центов. Вы заметили, что средний спрос на газету составляет 100 штук; при этом стандартное отклонение спроса составляет 20 штук. Какой заказ следует сделать, чтобы максимизировать прибыль?

Эту задачу мы будем решать в других частях курса; однако, обратите внимание, что даже для того, чтобы точно сформулировать задачу, необходимо обладать знаниями о спросе на газету. А для этого, в частности, нужно регистрировать не только все проданные экземпляры газеты, но и «необслуженный спрос». То есть на месте продавца газет

придется терпеливо стоять и записывать всех, кто пожелал приобрести, увы, закончившуюся газету. Будьте уверены, что даже в тех данных, которые вы вели собственноручно, вы найдете неточности — начиная с обыкновенных опечаток и заканчивая ошибками выборки.

Поэтому подготовка данных — дело, хотя и сравнительно простое, но необыкновенно важное. Как говорят американские бухгалтеры, есть правило FIFO (первый на приход — первый на уход) и правило LIFO (последний на приход — первый на уход); но правило GIGO — важнее (Garbage In — Garbage Out, то есть «мусор на входе — мусор на выходе»).

1.2. Выборки

Первая глава на тему инференционной статистики (статистики выводов) посвящена выборке. Понятно, что не всегда возможно вести речь о генеральной совокупности (популяции), дающей все возможные исходы или измерения, представляющие интерес; зачастую — прежде всего, ввиду ограниченности ресурсов — приходится вести речь о выборке, подмножестве совокупности. В этой главе мы побеседуем о том, почему в статистике имеют дело с выборками и каковы могут быть последствия их неправильного отбора.

Практически все статистические результаты основываются на измерении выборки, взятой из генеральной совокупности. Судьбоносные решения часто принимаются на основе информации, полученной из выборок. Например, рейтинги персонажей реалити-шоу формируются на основании собранной информации у небольшой выборки граждан, а на их основе делаются заключения в отношении телевизионной аудитории всей страны. Будущее вашего телевизионного шоу находится в руках этой небольшой группки граждан! Грамотный отбор выборки — это решающий шаг, влияющий на точность статистических выводов.

Большинство статистических исследований опирается на выборку, взятую из генеральной совокупности.

Почему бы не измерить всю генеральную совокупность вместо того, чтобы полагаться на выборку? В зависимости от исследования измерение генеральной совокупности может стоить слишком дорого или быть вообще невозможным. К примеру, если ученые захотят измерить продолжительность жизни одного из видов назойливых москитов, то — в обозримом будущем — им вряд ли удастся произвести наблюдение за всеми москитами в совокупности. Придется положиться на выборку генеральной совокупности москитов, измерить продолжительность их жизни, а затем сделать предположение относительно продолжительности жизни всей генеральной совокупности. В этом и состоит основополагающая идея статистического вывода! К сожалению, выполнить сказанное куда сложней, чем просто написать об этом.

Даже если бы мы могли измерить всю генеральную совокупность целиком, такой шаг мог бы оказаться совершенно бесполезным. Если выборка отобрана грамотно и анализ произведен правильно, мы можем сделать довольно точные выводы и оценку всей совокупности. Нет смысла выходить за пределы выборки и измерять все, что окажется в поле зрения. Измерение всей генеральной совокупности зачастую оборачивается впустую потраченным временем и деньгами — ресурсами весьма дефицитными.

Случайная выборка

Термин случайная выборка относится к процедуре отбора, при котором все представители совокупности имеют равные шансы быть отобранными. Цель случайной выборки — удостовериться, что финальная выборка, подлежащая измерению, является репрезентативной в отношении всей совокупности, из которой она была взята. Если же это не так, то мы имеем дело с выборкой с пристрастием, измерение которой может привести к неверным результатам. Грамотный отбор выборки является решающим для точности статистического анализа.

Существует несколько способов отбора случайной выборки. Для их демонстрации я воспользуюсь следующим примером. Допустим, вы хотите произвести опрос посетителей магазина на предмет их мнения относительно внешнего вида человека, фотографию которого вы им показываете. Каким образом вы будете выбирать тех, кого будете опрашивать? (Примечание: наш магазин устроен так, что в день его посещают ровно 1000 человек).

Простая случайная выборка: способы отбора

Простая случайная выборка — это выборка, в которой все представители совокупности имеют равные шансы быть отобранными. Но проще сказать, чем сделать. В примере с универмагом мы можем случайным образом выбирать людей и спрашивать их мнение. Но в нашем отборе могут быть пристрастия. Например, если я увижу некоего типа угрожающего вида с татуировкой «Смерть математикам!», вряд ли я выберу его, чтобы узнать, что он думает. Разумеется, с точки зрения анализа, я поступлю необъективно.

Допустим, я могу избавить себя от выборки с пристрастием, тогда примером простой случайной выборки будет слудующее:

Каждый «X» — это покупатель, а каждый «X», обведенный кругом, —покупатель, вошедший в мою выборку.

Есть и другие **способы отбора** простой случайной выборки для опроса. Я мог бы случайно отобрать персонажей *с помощью таблицы случайных чисел*. Ниже показана часть такой таблицы.

57245	39666	18545	50534	57654	25519	35477	71309	12212	98911
42726	58321	59267	72742	53968	63679	54095	56563	09820	86291
82768	32694	62828	19097	09877	32093	23518	08654	64815	19894
97742	58918	33317	34192	06286	39824	74264	01941	95810	26247
48332	38634	20510	09198	56256	04431	22753	20944	95311	29515
26700	40484	28341	25428	08806	98858	04816	16317	94928	05512
66156	16407	57395	86230	47495	13908	97015	58225	82255	01956
64062	10061	01923	29260	32771	71002	58132	58646	69089	63694
24713	95591	26970	37647	26282	89759	69034	55281	64853	50837
90417	18344	22436	77006	87841	94322	45526	38145	86554	42733

Положим, наша генеральная совокупность состоит из 1000 человек, из которых нам необходимо отобрать выборку в 100. Пронумеруем вошедших от 0 до 999. В соответствии с таблицей случайных чисел будет отобран посетитель 572, затем 427 и так далее, пока не будут отобраны 100 человек (поскольку таблица велика, а нам нужно отобрать всего 100, то мы просто отбрасываем последние 2 цифры — очевидно, что число, образованное первыми 3-мя цифрами случайного числа, тоже является случайным).

Используя такую методику, мы произведем совершенно случайную выборку.

Случайные числа также можно сгенерировать *с помощью функции СЛЧИС() программы Excel*. Ячейка A1 содержит формулу = СЛЧИС(), которая предоставляет собой случайное число от 0 до 1. Умножив это число на 1000 и взяв целую часть, мы получим случайное целое число, равномерно распределенное на промежутке от 0 до 999..

Систематическая выборка

Один из способов избежать пристрастности при случайном отборе людей — это использование систематической выборки. Эта методика подразумевает отбор каждого **k** члена совокупности, который и будет представлен в выборке. Значение **k** зависит от размера выборки и генеральной совокупности. В примере с универмагом при размере совокупности в 1000 посетителей и выборке величиной 100 **k = 10**. Из списка всей совокупности я буду отбирать для выборки каждого десятого покупателя. В целом, если **N** = размер генеральной совокупности, **n** = размер выборки, тогда: **k=N/n**.

Мы также можем применить эту методику к примеру с универмагом. Каждому третьему посетителю универмага будет задан вопрос по поводу фотографии, даже если этот посетитель будет обладателем угрожающей татуировки. Мы снова можем принять те же обозначения: «X» — посетитель, а «X», обведенный кругом, — посетитель, вошедший в выборку.

Преимуществом использования систематической выборки является простота ее поведения по сравнению с простой случайной выборкой — такая выборка часто требует меньших затрат времени и средств. **Недостаток** — опасность отбора пристрастной выборки, если в совокупности прослеживается поведение, сопоставимое со значением k. Например, предположим, что я провожу опрос в студенческом городке, узнавая у студентов, сколько часов в неделю они посвящают учебе. Для сбора данных я выбираю январь, май, июнь и декабрь. Поскольку в большинстве учебных заведений сессия приходится именно на это время, мы получим существенно искаженные данные.

Групповая выборка

Если генеральную совокупность можно разделить на группы, то простая случайная выборка может быть произведена из этих групп для формирования финального варианта выборки. В примере с учебным заведением (университетом, институтом, колледжем) в качестве групп могут выступать учебные группы, которые мы будем отбирать случайным образом для участия в опросе. В каждой из выбранных групп все студенты будут включены в выборку.

Любой статистик больше всего беспокоится по поводу **ошибок выборки**, которые происходят, когда измерение выборки отличается от измерения совокупности. Поскольку генеральная совокупность целиком измеряется крайне редко, невозможно совершенно точно вычислить ошибку выборки. И все-таки с помощью статистики вывода мы научимся определять вероятности некоторого количества ошибок выборки.

Ошибки выборки случаются тогда, когда мы производим неудачный отбор выборки, не соответствующей своей генеральной совокупности. Например, если большинству посетителей универмага очень понравилась внешность человека на фотографии, но нам случилось выбрать тех, кто не способен ценить внешний облик, тогда, возможно, мы придем к неправильному выводу.

К ошибкам выборки следует быть готовым: они являются своего рода платой за то, что нам не приходится обрабатывать всю совокупность целиком. Одним из способов уменьшения вероятности ошибки выборки статистического исследования является увеличение размера выборки. В целом, чем больше размер выборки, тем меньше вероятность ошибки. Если вы увеличите размер выборки до размера генеральной совокупности, то ошибка выборки будет равна нулю. Но таким образом вы лишитесь всех достоинств выборки.

Драматичные примеры использования ошибочных выборочных методик

Большинство примеров взяты из американского опыта; это, разумеется, не потому, что там делают наибольшее количество ошибок, — скорее, потому, что именно в США статистика имеет наибольший вес как наука.

Выборочные методики широко используются в политике. Но используются они не всегда грамотно.

Одна из самых больших неудач при осуществлении выборки произошла во время президентской гонки 1936 года, когда Литературный Дайджест предсказал, что Альф Лэндон одержит победу над Франклином Рузвельтом на выборах Президента США. Даже если вы не очень сильны по части истории, вы наверняка поняли, что кое-кто после выборов оказался в весьма неприятном положении. Литературный Дайджест отобрал выборку из телефонных книг и регистрационных книг автомобилистов. Но проблема состояла в том, что в 1936 году владельцами телефонов и автомобилей являлись в основном состоятельные республиканцы, выборка которых не являлась репрезентативной для всей совокупности избирателей. Ошибка стоила журналу жизни — читатели-республиканцы не простили его, ведь журнал предрекал уверенную победу их кандидату.

Другая политическая ошибка подобного рода была допущена в 1948 году, когда во время президентской гонки Институт Гэллапа предсказал, что Томас Дьюи одержит победу над Гарри Трумэном. Неудача Института Гэллапа состояла в том, что в их выборке оказалось большое количество неопределившихся избирателей. Было сделано неверное предположение о том, что эти избиратели являются репрезентативной выборкой определившихся избирателей, поддерживающих Дьюи. Трумэн без труда выиграл на выборах, набрав 303 голоса против 189 голосов у Дьюи.

Кстати, квинтэссенцией неправильных выборок можно считать следующий анекдот: «По результатам **он-лайн** опроса выяснилось, что 100% жителей Российской Федерации хотя бы раз в жизни пользовались Интернетом».

Как видите, для инференционной статистики грамотный отбор выборки является решающим шагом. Даже большой размер выборки не способен скрыть ошибки отбора выборки, не являющиеся репрезентативной в отношении совокупности в целом. История показала, что большие размеры выборок вовсе не обеспечивают точности. Например, Институт Гэллапа предсказал, что Никсон получит 43% голосов на президентских выборах 1968 года, а он получил 42,9%. В данном случае Институт Гэллапа опирался на размер выборки всего в 2 тысячи человек, в то время как Литературный Дайджест опросил 2 миллиона человек — огромная выборка по любым меркам.

1.3. Проверка гипотез

Статистики любят делать предположения относительно параметра совокупности, отбирать выборку из этой совокупности, измерять выборку и объявлять, было ли первоначальное предположение подтверждено выборкой. Это — в двух словах — и есть алгоритм проверки гипотезы.

В мире статистики **гипотезой** называется предположение о параметре генеральной совокупности. Примерами гипотез являются такие предположения:

- взрослый человек в среднем выпивает 1,7 чашки кофе ежедневно;
- 12% студентов после окончания университета сразу отправляются в аспирантуру;

- не более 2% товаров, продаваемых нашими заказчиками, окажутся с дефектом.

В каждом случае мы сделали предположение относительно совокупности, которое может оказаться истинным или ложным. Цель проверки гипотезы — сделать **статистический вывод** о принятии или отклонении этих предположений.

Основная и альтернативная гипотезы

Каждая проверка гипотезы подразумевает наличие основной (нулевой) и альтернативной гипотез. **Основная (нулевая) гипотеза, обозначаемая H_0**, неизменна и высказывает предположение, что среднее по совокупности «равно», «больше или равно» или «меньше или равно» определенному значению («=», «>=» и «=<» соответственно). Основная гипотеза считается истинной, если нет явного доказательства противоположного.

Альтернативная гипотеза, обозначаемая H_1, содержит утверждение, противоположное утверждению основной гипотезы, и считается верной, если основная гипотеза оказывается ложной (аккуратные статистики говорят «гипотеза опровергнута (или не принята)», или, в противном случае «гипотеза не может быть опровергнута» — именно так, с математической точки зрения, следует приводить результаты). Альтернативная гипотеза всегда утверждает, что среднее по генеральной совокупности «сторого больше», «сторого меньше» или «не равно» определенному значению.

Обратите внимание, что альтернативная гипотеза никогда не использует несторгие неравенства.

Формулировка основной и альтернативной гипотез

Будьте предельно внимательны при формулировке основной и альтернативной гипотез. Ваш выбор будет зависеть от характера проверки и мотивации человека, ее проводящего.

Если целью является проверка того, что среднее по совокупности равно определенному значению, назначьте такое утверждение основной гипотезой.

Часто проверка гипотезы проводится исследователями с целью доказать, что их открытие существенно улучшает существующие продукты или процедуры. Например, если я изобрел мяч для гольфа и утверждаю, что он летит дальше обычных мячей более чем на 20 метров, моя гипотеза будет выглядеть следующим образом:

H_0: μ <= 20
H_1: μ > 20

Обратите внимание, что мы использовали альтернативную гипотезу для формулировки утверждения, которое хотим доказать статистически, с целью заработать себе состояние на продаже этих мячей отчаявшимся игрокам в гольф. Поэтому *альтернативная гипотеза также носит название исследовательской*, поскольку представляет позицию, которую хочет утвердить и закрепить исследователь.

Двусторонняя проверка гипотезы

Двусторонняя проверка гипотезы используется в случае, если альтернативная гипотеза сформулирована, например, как H=20.

Процедура выглядит просто:

- отобрать выборку размером **n** и вычислить выборочный показатель – в данном случае среднее по выборке;
- отложить среднее по выборке на оси **x** кривой выборочного распределения;
- если среднее по выборке оказывается в пределах белой области, мы не отклоняем H_0, то есть, у нас нет достаточных доказательств для поддержки H_1, альтернативной гипотезы, утверждающей, что среднее по совокупности не равно 20 метрам;
- если среднее по выборке попадает в одну из заштрихованных областей, называемых *областями отклонения гипотезы*, мы отклоняем H_0, то есть мы обладаем необходимым доказательством для поддержки H_1 и убеждены, что истинное среднее по совокупности не равняется 20 метров.

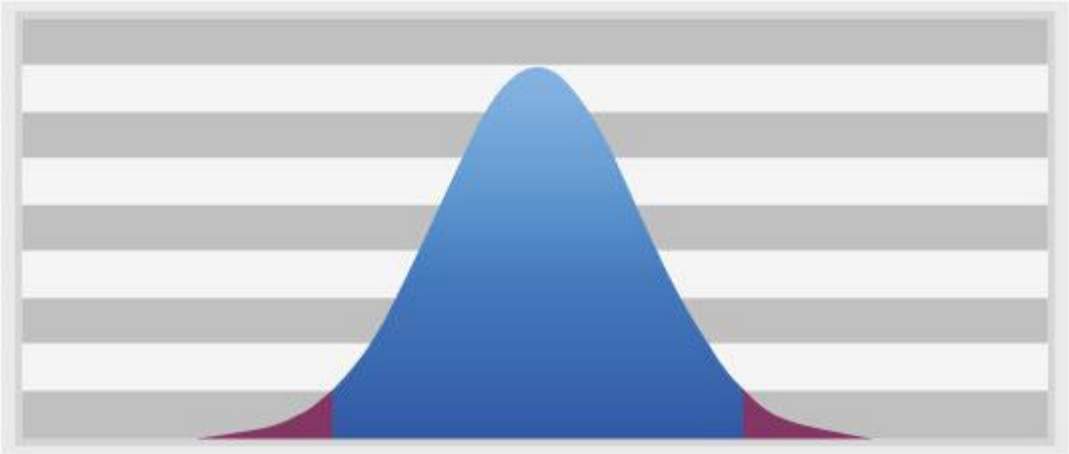

Рис. 1. Двусторонняя проверка гипотезы

Поскольку на рисунке две закрашенные области, мы имеем дело с двусторонней проверкой гипотезы.

Односторонняя проверка гипотезы

Односторонняя проверка гипотезы относится к альтернативной гипотезе, сформированной в виде «>=» или «<=». Так, пример с мячами для гольфа как раз требует односторонней проверки, поскольку альтернативная гипотеза выглядит как H_1 μ>20. График такой проверки показан на рисунке.

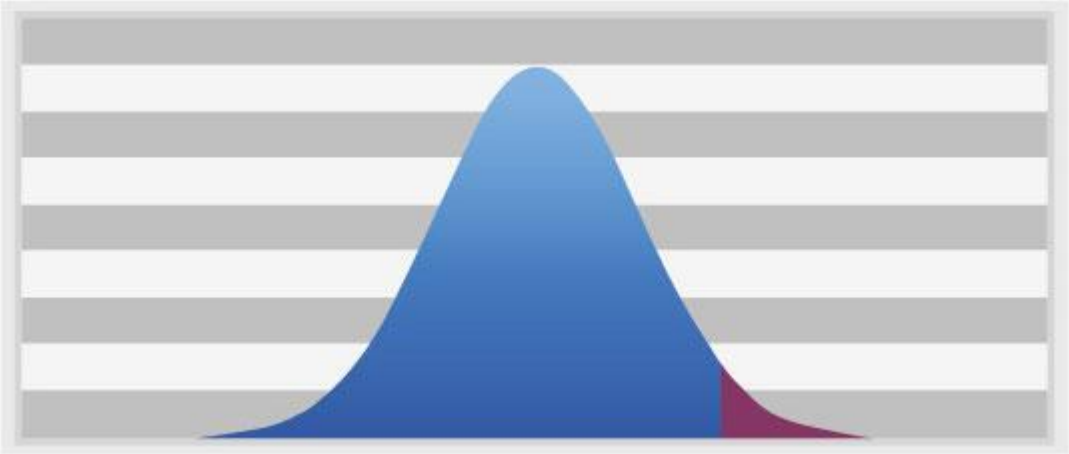

Рис. 2. Односторонняя проверка гипотезы

На этом графике мы видим лишь одну область отклонения — заштрихованную область правого «хвоста» распределения. Следуем той же процедуре, которую использовали для двусторонней проверки, и откладываем среднее по выборке, которое представляет собой среднее увеличение расстояния от площадки для первого удара с помощью моего нового мяча для гольфа.

Существует **два возможных сценария:**

- если среднее по выборке попадает в рамки незаштрихованной области, мы не отклоняем H_0, то есть у нас нет достаточных доказательств для поддержки H_1, альтернативной гипотезы, утверждающей, что изобретенный мяч увеличивает расстояние от площадки не более чем на 20 метров;
- если среднее по выборке оказывается в пределах области отклонения, мы отвергаем H_0, то есть мы обладаем достаточными доказательствами для поддержки H_1, утверждающей, что новый мяч увеличит расстояние от площадки более чем на 20 метров. Пора уходить на пенсию и зарабатывать на новом изобретении!

1.4. Ошибки первого и второго рода

Помните, что цель проверки гипотезы состоит в подтверждении утверждения относительно совокупности на основе одной выборки. Поскольку мы полагаемся на выборку, то подвергаем себя риску, что наши выводы о совокупности могут оказаться ошибочными.

Используя пример с мячами для гольфа, положим, что моя выборка попадает в область «Отклонить H_0» на последнем рисунке. То есть в соответствии с выборкой, мой мяч увеличит дистанцию более чем на 20 метров. А что, если истинное среднее по совокупности на самом деле значительно меньше 20 метров? Это может произойти в результате ошибки выборки. Тип ошибки, когда мы отклоняем H_0, а на самом деле это

гипотеза является истинной, называется **ошибкой первого рода**. Вероятность совершения ошибки первого рода определяется уровнем значимости и обозначается **α**.

При проверке гипотезы может произойти и другого рода ошибка. Предположим выборка с мячом для гольфа оказалась в пределах области «Не отклонять H_0» на рис. 2. То есть в соответствии с выборкой, мяч увеличивает дистанцию не более чем на 20 метров. Но что, если истинное среднее по совокупности на самом деле больше 20 метров? Тип ошибки, когда мы не отклоняем H_0, а на самом деле она является ложной, называется **ошибкой второго рода**. Вероятность совершения ошибки второго рода называется мощностью гипотезы и обозначается **β**.

В табл. 1 представлены оба типа ошибок проверки гипотезы.

Таблица 1. Два типа ошибок проверки гипотезы

	H_0 истинна	**H_0 ошибочна**
Отклонить H_0	Ошибка первого рода P [ошибка первого рода] = α	Правильный исход
Не отклонять H_0	Правильный исход	Ошибка второго рода P [ошибка второго рода] = β

Как правило, при проверке гипотезы мы определяем *уровень значимости* **α**, который находится в пределах от 0.01 до 0.10, и происходит это до отбора выборки.

Что бы вам было проще запомнить, можно думать об ошибках первого и второго рода следующим образом: нулевая гипотеза — это гипотеза о невиновности обвиняемого (следуя презумпции невиновности). Таким образом, (потенциальная) ошибка, совершить которую вы боитесь больше всего, — это признать невиновного виновным. Соответственно, именно ошибками первого рода больше всего озабочена статистика.

1.5. Использование шкалы исходной переменной

В этом разделе мы определим область отклонения с помощью **шкалы исходной переменной**, которой в нашем примере является количество дней. Для вычисления верхней и нижней границ области отклонения используем следуюшие уравнения.

Границы области отклонения = μ_{H0} **+/- z ×** σ_x, где μ_{H0} = среднее по совокупности, принятой основной гипотезой, а σ_x — стандартное отклонение по выборке.

Для нашего примера с заработной платой (см.Приложение 1):

Верхняя граница = 16732

Нижняя граница = 63819

Поскольку наше среднее по подвыборке №1 из 20 элементов равно 36,653 (округленно), по этой подвыборке мы попадаем в область «Не отклонять H_0» — как на рис. 1. Отсюда делаем вывод, что разница между 36,653 и 40,000 — это исключительно дело случая, и у нас есть «подтверждение», что среднее по совокупности равняется 40,000 (в действительности, лучше считать, что мы не можем опровергнуть, что среднее по совокупности равно 40,000).

Использование стандартизованной нормальной шкалы

Мы можем получить то же самое заключение, установив границы области отклонения с помощью нормальной шкалы. Для этого вычисляем z-распределение, соответствующее выборочному среднему, как показано ниже:

$$Z = (x - \mu_{H0})/ \sigma_x.$$

Затем можно пользоваться таблицами стандартного нормального распределения или функцией Excel НОРМСТОБР().

Ранее мы с вами применяли статистику вывода для заключений в отношении одной, двух или более средних и долей по совокупности.

Теперь речь пойдет о том, как переменные могут быть связаны друг с другом. С помощью корреляции и линейной регрессии мы сможем, во-первых, определить, существует ли связь между двумя переменными, а во-вторых, описать природу этой связи в математических терминах.

1.6. Независимые и зависимые переменные

Допустим, я хочу определить, существует ли связь между количеством часов, посвященных студентом изучению статистики, и финальной экзаменационной оценкой. В табл. 2 представлены выборочные данные о 6-ти случайным образом отобранных студентах.

Таблица 2. Данные для экзамена по статистике

Количество часов учебы	Экзаменационная оценка (из 100 баллов)
3	86
5	95
4	92
4	83
2	78
3	82

Достаточно очевидно, что количество часов положительно сказывается на финальной оценке. Переменная «Количество часов учебы» считается независимой переменной (x), поскольку она приводит к наблюдаемой вариации переменной «Экзаменационная оценка», которая в нашем случае считается зависимой переменной (y). Данные из таблицы 2 считаются упорядоченными парами (x,y) значений, такими как (3.86) и (5.95).

«Причинная связь» между зависимыми и независимыми переменными существует только в одном направлении:

Независимая переменная (x) -> Зависимая переменная (y)

В обратном направлении эта связь не работает. Вряд ли можно представить, что оценка может быть причиной более продолжительного изучения предмета студентом.

Другие примеры зависимых и независимых переменных представлены в табл. 3.

Таблица 3. Примеры зависимых и независимых переменных

Независимая переменная	Зависимая переменная
Размер телевизора	Цена телевизора
Уровень рекламы	Объем продаж
Размер оплаты игроков	Число побед (с последним, разумеется, можно спорить)

Далее речь пойдет о связи между переменными **x** и **y** с использованием статистики вывода.

1.7. Корреляция

Корреляция измеряет мощность и направление связи между **x** и **y**. На рис. 3 представлены различные типы корреляции в виде графиков рассеяния упорядоченных пар (x,y). По традиции переменная **x** размещается на горизонтальной оси, а **y** — на вертикальной.

График A на рис. 3 является примером положительной линейной корреляции: при увеличении **x** также увеличивается **y**, причем линейно. **График B** показывает нам пример отрицательной линейной корреляции, на котором **y** при увеличении **x** линейно уменьшается. **На графике C** мы видим отсутствие корреляции между **x** и **y**. Эти переменные никоим образом не влияют друг на друга.

Наконец, **график D** — это пример нелинейных отношений между переменными. По мере увеличения **x**, наш **y** сначала уменьшается, а потом меняет направление и увеличивается.

Теперь поговорим о линейных взаимосвязях между зависимой и независимой переменными.

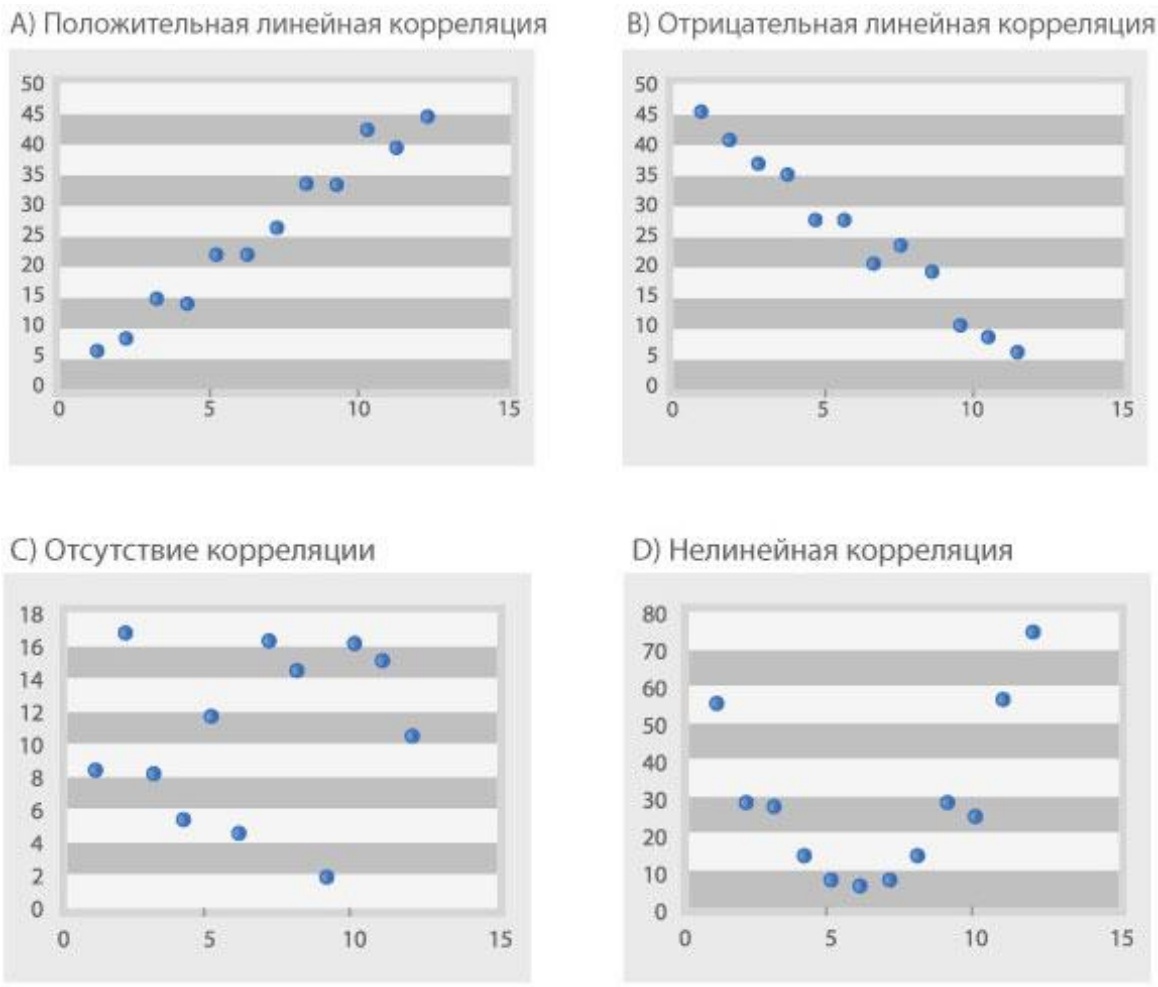

Рис. 3 Различные типы корреляции

Коэффициент корреляции

Коэффициент корреляции **r** демонстрирует как силу, так и направление связи между независимой и зависимой переменными. Значения r могут находиться в диапазоне от -1.0 до +1.0. Когда **r** положителен, связь между x и y также является положительной (график A на рис. 3), а когда значение r отрицательно, связь также отрицательна (график B). Коэффициент корреляции, близкий к нулевому значению, свидетельствует о том, что между **x** и **y** связи не существует (график C).

Сила связи между **x** и **y** определяется близостью коэффициента корреляции к -1.0 или +1.0. Обратите внимание на рис. 4.

График A показывает идеальную положительную корреляцию между **x** и **y** при **r** = + 1.0. **График B** — идеальная отрицательная корреляция между **x** и **y** при **r** = - 1.0. **Графики C и D** — примеры более слабых связей между зависимой и независимой переменными.

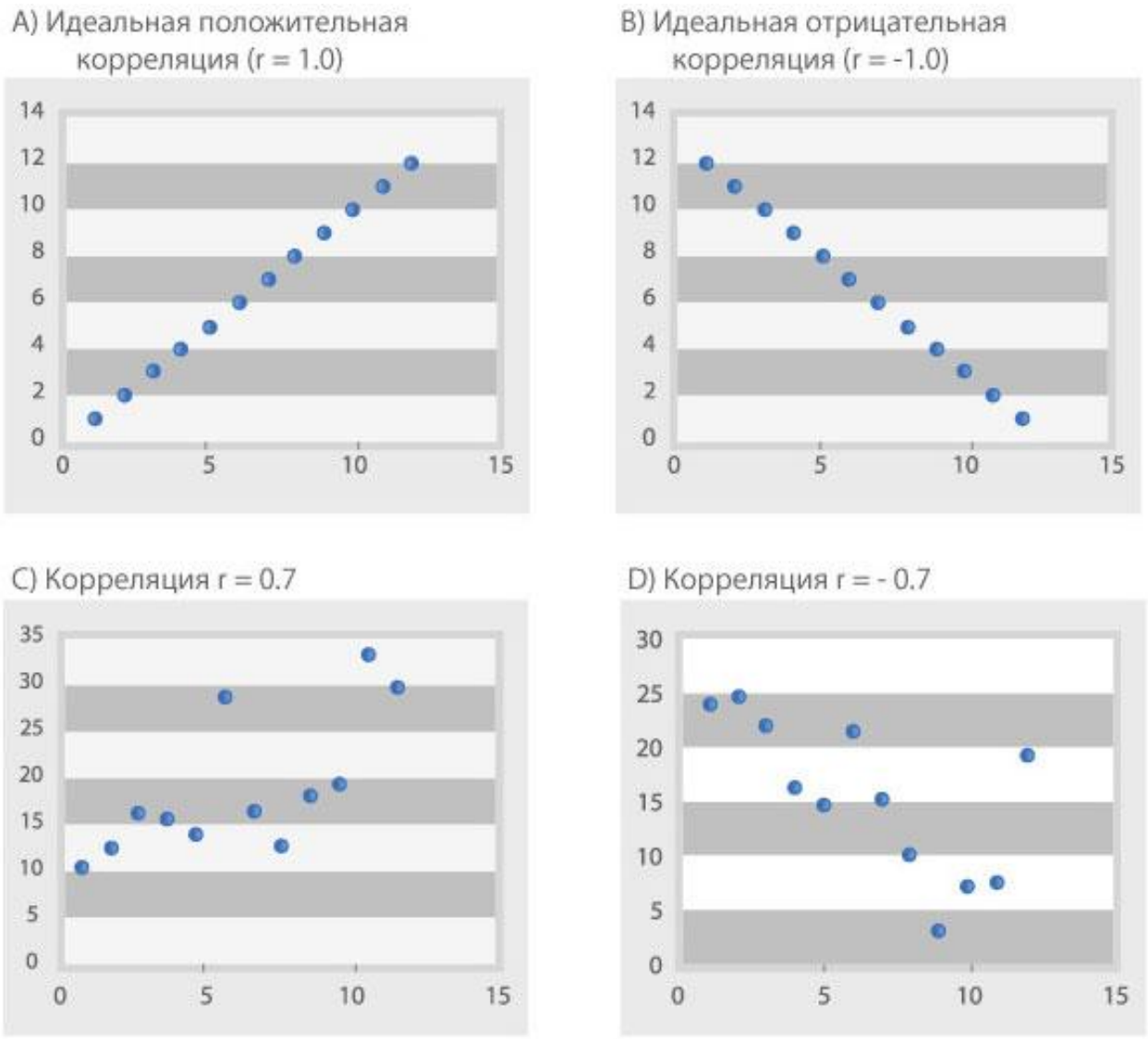

Рис. 4.

Развернутая формула коэффициента корреляции

Приведем (развернутую) формулу коэффициента корреляции:

$$r = \frac{n \times \sum x \times y - (\sum x)(\sum y)}{\sqrt{[n \times \sum x^2 - (\sum x)^2] \times [n \times \sum y^2 - (\sum y)^2]}}$$

Табл. 4 поможет нам разбить формулу на несколько несложных вычислений.

Таблица 4. Формула коэффициента корреляции

Часы изучения X	Экзамен У	Произведение x × y	x^2	y^2
3	86	258	9	7396
5	95	368	25	8464
4	92	475	16	9025
4	83	332	16	6889
2	78	156	4	6084
3	82	246	9	6724
Итого:				
21	516	1835	79	44 582

Используя эти значения и **n = 6** (число упорядоченных пар), получаем:

$$r = \frac{6 \times 1835 - 21 \times 516}{\sqrt{[6 \times 79 - 21^2] \times [6 \times 44582 - 516^2]}} = \frac{174}{\sqrt{33 \times 1236}} = 0.862$$

Как видите, между числом часов, посвященных изучению предмета, и экзаменационной оценкой существует весьма сильная положительная корреляция. Преподаватели будут весьма рады узнать об этом.

Какова выгода устанавливать связь между подобными переменными? Отличный вопрос. Если обнаруживается, что связь существует, мы можем предугадать экзаменационные результаты на основе определенного количества часов, посвященных изучению предмета. Проще говоря, чем сильнее связь, тем точнее будет наше предсказание. Мы научимся делать подобные предварительные оценки, когда перейдем к теме линейной регрессии.

Программа Excel может выполнить за вас всю эту работу с помощью функции КОРРЕЛ со следующими характеристиками:

КОРРЕЛ (массив 1; массив 2), где:

массив 1 = диапазон данных для первой переменной, массив 2 = диапазон данных для второй переменной.

1.8. Линейная регрессия

Метод линейной регрессии позволяет нам описывать прямую линию, максимально соответствующую ряду упорядоченных пар (x,y). Уравнение для прямой линии, известное как линейное уравнение, представлено ниже:

$$y = a + b x$$

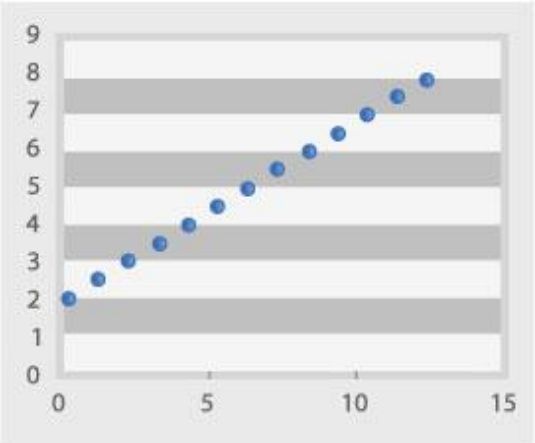

Рис. 5.

На рисунке 5 показана линия, описанная уравнением y = 2+0.5x. Отрезок на **оси y** — это точка пересечения линией **оси y**; в нашем случае **a = 2**. Наклон линии **b**, отношение подъема линии к длине линии, имеет значение 0.5. Положительный наклон означает, что линия поднимается слева направо. Если **b = 0**, линия горизонтальна, и это значит, что между зависимой и независимой переменными нет никакой связи. Иными словами, изменение значения **x** не влияет на значение **y**.

Следующий шаг — определить линейное уравнение, максимально соответствующее набору упорядоченных пар.

Метод наименьших квадратов

Метод наименьших квадратов (МНК, или OLS — Ordinary Least Squares) — это математическая процедура составления линейного уравнения, максимально соответствующего набору упорядоченных пар, путем нахождения значений для **a** и **b**, коэффициентов в уравнении прямой. **Цель метода** наименьших квадратов состоит в минимализации общей квадратичной ошибки между наблюдавшимися и предсказанными значениями. Если для каждой точки мы определяем ошибку **y**, то можем построить линию регрессии так, чтобы минимизировать следующую сумму:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

где n — число упорядоченных пар вокруг линии, максимально соответствующей данным.

Таким образом, линия *регрессии* минимизирует общую квадратичную ошибку.

Рассмотрим, как вычисляются коэффициенты регрессии, на следующем примере.

Пример вычисления коэффициента регрессии

Представьте, что вы — руководитель компании, одним из основных направлений деятельности которой является срочная (курьерская) доставка. В последнее время, погнавшись за скоростью доставки, вы перестроили систему мотивации водителей-курьеров таким образом, что она стала в значительной мере зависеть от скорости доставки. Неприятным побочным эффектом смены системы мотивации стало увеличение количества ДТП, в которые попадают водители. И хотя вы сознаете, что нежелательный результат мог быть связан с целым рядом возможных воздействий, вы, в первую очередь, хотите проверить, верна ли ваша гипотеза об увеличении количества ДТП или ваши подозрения не обоснованы статистически.

Данные о количестве ДТП

Месяц	Количество ДТП	Месяц	Количество ДТП
1	8	6	13
2	6	7	9
3	10	8	11
4	6	9	15
5	10	10	17

Поскольку своей целью мы поставили задачу узнать, увеличивается ли со временем число ДТП, «Месяц» будет независимой переменной, а «Количество ДТП» — зависимой.

Для определения уравнения регрессии мы будем пользоваться Excel — соответствующая функция называется ЛИНЕЙН() и имеет формат ЛИНЕЙН (Известные_Y, Известные_X, A, Статистика).

Здесь Известные_Y — вектор значений Y, Известные_X — таблица значений X (вектор в случае одной переменной), A — может принимать значение 0 или 1 и определяет, должна ли линия регрессии выходить из 0, Статистика — может принимать значение 0 или 1 и определяет, нужно ли выводить дополнительные данные по регрессии.

Ниже приведена таблица результатов регрессии.

Наклон	0,975758	5,133333	Пересечение
Точность определения наклона	0,246034	1,526599	Точность определения пересечения
r^2 (Квадрат коэффициента корреляции)	0,662856	2,234712	S_e — точность прогноза
F-тест	15,72876	8	Число степеней свободы
SS регрессии	78,54848	39,95152	SS остатков

Кривая эффекта для нашего примера будет определяться следующим уравнением:

$$y = 5.\ 13 + 0.976\ x$$

Поскольку наше уравнение имеет положительный наклон +0.976, мы имеем доказательства того, что **число ДТП со временем увеличивается со средней скоростью 1 в месяц**.

На рис. 6 представлена линия регрессии — вместе с наблюдавшимися значениями.

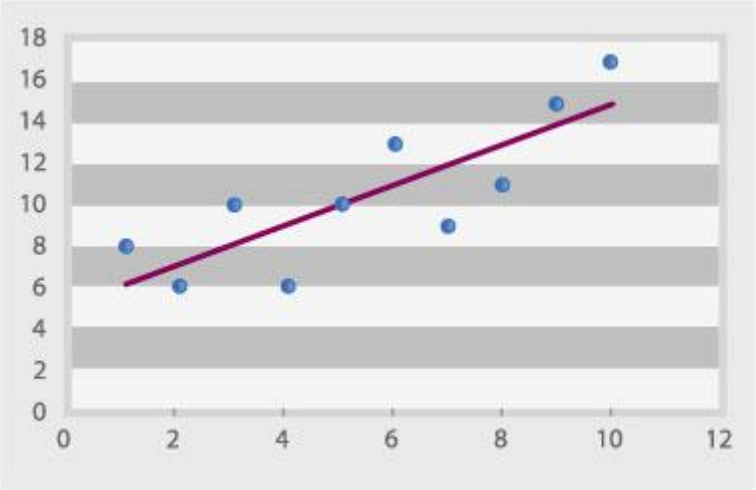

Рис. 6.

Таким образом, наше ожидание в отношении числа ДТП в течение следующего периода (месяца 11) будет вычисляться так:

$$y = 5.13+0.976x=5.13 +0.976(11) =15.87 \sim 16.$$

Может, пора опять менять систему мотивации?

Доверительный интервал кривой эффекта

Насколько точны мои ожидания в отношении числа ДТП на определенный месяц? Чтобы ответить на этот вопрос, нам необходимо найти *оценку стандартной ошибки* S_e. К счастью, Excel уже сделал это за нас: S_e =2.2347~2.23

Теперь мы можем вычислить доверительный интервал среднего **y** вокруг определенного значения **x**. В Месяце 8 (x = 8) произошло 11 ДТП (y = 11). Из линии регрессии ожидаем, что:

$$y = 5.13 +0.976\ x =5.13 +0.976(8) =12.9 \text{ ДТП}.$$

Формулу для доверительного интервала заинтересованный слушатель сможет найти в литературе; мы же приведем только окончательный результат:

Число ДТП (95% доверительный интервал) для месяца 8 находится в пределах от 10.74 до 15.06.

Проверка наклона линии регрессии

Вспомним, что если наклон кривой эффекта **b** равняется нулю, между переменными **x** и **y** *нет* никакой взаимосвязи. В нашем примере с числом ДТП мы вычислили, что наклон кривой эффекта равен 0.976. Но поскольку этот результат основан на выборке наблюдений, нам необходимо проверить, действительно ли 0.976 находится довольно далеко от нуля, чтобы подтвердить, что между двумя переменными действительно существует связь. Если это наклон фактической совокупности, тогда формулируем гипотезы так:

H_0: $\beta = 0$
H_1: $\beta <> 0$

Если мы отклоним основную гипотезу, то сможем сделать заключение, что на основе этой выборки между зависимой и независимой переменными действительно существует связь.

Проверим это при **a = 0.01**.

Проверка этой гипотезы потребует определить *стандартную ошибку наклона* S_b. В таблице вывода Excel стандартная ошибка расположена непосредственно под вычисленным коэффициентом.

Критерий значимости для данной гипотезы будет рассчитываться так:

$$t = \frac{b - \beta_{H_b}}{s_b} = \frac{0.976}{0.246} = 3.966$$

Критическое значение t= t_c найдем из t-распределения Стьюдента при n – 2 = 10 - 2 = 8 степенях свободы. При двусторонней проверке a = 0.01 t_c = 3.355 в соответствии с таблицами или функцией Excel. Поскольку t > t_c, мы отклоняем основную гипотезу и заключаем, что между месяцем и числом ДТП действительно существует связь.

Допущения для линейной регрессии

Чтобы все эти результаты были действительными, нам необходимо убедиться, что не нарушаются допущения линейной регрессии.

- Существует линейная связь между независимой и зависимой переменными.

- Остатки (индивидуальные различия между данными и линией, определяемой уравнением регрессии) являются независимыми друг от друга.
- Наблюдаемые значения **y** являются нормально распределенными вокруг ожидаемого значения (или, формулируя на языке остатков, «остатки являются нормально распределенными со средним, равным 0»).
- Вариация y вокруг кривой эффекта равняется всем значениям **x**.

Методики для проверки этих допущений не входят в рассмотрение этого курса.

Линейная регрессия на несколько переменных (Множественная регрессия)

Линейная (простая) регрессия ограничивается рассмотрением связи между зависимой переменной и только одной независимой переменной. Если в связи присутствует более одной независимой переменной, тогда нам необходимо обратиться к множественной регрессии. Уравнение для такой регрессии выглядит следующим образом:

$$y = a+b_1x, +b_2x_2 + ...+ b_nx_n$$

Как вы понимаете, здесь все гораздо сложнее. Возникает целый ряд проблем; некоторые из них мы просто отметим, на некоторых — наиболее важных и интуитивно понятных — остановимся подробнее.

Чем хороша одномерная регрессия? Ее график можно изобразить на плоскости. С увеличением числа переменных такая возможность, увы, пропадает. Уже в случае двух независимых переменных изображение графика представляет некоторые сложности. Если же вы научитесь строить интуитивно понятные графики зависимостей от 3 или более независимых переменных — пожалуйста, сообщите мне. Смотря по ситуации, я должен буду передать это сообщение дальше — либо в Нобелевский комитет, либо в соответствующую больницу.

Проблемы, возникающие в случае множественной регрессии

Таким образом, **первая проблема**: сложности с интерпретацией и наглядностью.

Далее. Когда модель имеет всего одну независимую переменную, сама по себе регрессия становится достаточно очевидной. Все меняется, когда приходят они — другие переменные.

Например, несколько лет назад в Соединенных Штатах провели опрос на тему «Количество КПК на 1000 человек населения в зависимости от возраста» (см. Приложение 2; данные условные!). Построив регрессию количества КПК на 1000 человек на возраст, исследователи получили результат, прямо противоположный тому, который рассчитывали получить: с возрастом количество КПК не уменьшалось, а возрастало!

Слава богу, организаторы опроса догадались включить и другие переменные. Оказалось, что количество КПК больше зависит от заработной платы; а так как последняя, в свою

очередь, зависит от возраста (по крайней мере, в Соединенных Штатах), то эта зависимость была перенесена и на количество КПК.

Таким образом, **проблема вторая** — мультиколлинеарность и, соответственно, одно из ее негативных проявлений — «пришпоренная» корреляция (то есть корреляция зависимой переменной с некой независимой переменной, вызванная, в свою очередь, корреляцией этой независимой переменной с другой независимой переменной).

Как бы хотелось, чтобы этим и исчерпывался список проблем с многомерной регрессией... Но — увы и ах — это далеко не все. Или, если пользоваться другой терминологией, даже близко не все.

Представьте себе, что вы ищете независимую переменную, которая лучше других предсказывает интересующую вас переменную — например, движение индекса РТС. Все шансы за то, что, перерыв изрядное количество данных, вы, в конце концов, воскликнете: «Эврика! Движение индекса РТС за последний год почти точно повторяет движение дневной температуры в Урюпинске!» (или в Кологриве, Конотопе, Крыжополе — просто подставьте ваш любимый город). Это, кстати, еще ничего — хотя вам будет несложно основывать торговую модель на движениях температуры, вам, по крайней мере, не придет в голову влиять на индекс с помощью установки нагревательного или охлаждающего оборудования. А представьте несколько другую ситуацию: вы обнаруживаете, что движение индекса РТС отслеживает движение цен на свежесваренных раков на Даниловском рынке.

Что дальше? Да все очень просто: вы продаете как можно больше акций «в короткую» (то есть совершаете действие, которое позволит вам получить прибыль при снижении цен), и каждый час посылаете на Даниловский рынок по небольшому грузовичку, нагруженному вашим любимым сортом раков. Последствия понятны: цена на раков падает, а за ней — и индекс РТС.

Итого, **проблема третья** — «раскапывание» данных (впрочем, она характерна и для одномерного случая).

Теперь давайте подумаем: как будет влиять добавление новых переменных на $\mathbf{R}^2$? Ответ сравнительно несложен: в худшем случае, добавление еще одной переменной *никак* не скажется на $\mathbf{R}^2$. То есть, самое плохое, что может сделать регрессия — просто не учитывать переменную, поставив при ней коэффициент 0. В особо тяжелых случаях, отдельные индивидуумы рассматривают такое количество переменных, что $\mathbf{R}^2$ становится равным единице.
Представьте, что у вас есть 365 наблюдений и 364 линейнонезависимых переменных. С их помощью (не забывайте про интерсепт — это, можно сказать, еще одна переменная) мы можем *с абсолютной точностью* описать все 365 наблюдений. Следовательно, $\mathbf{R}^2$ будет равен единице.
Тогда проблема четвертая — «Игра в $\mathbf{R}^2$».

Предположения множественной регрессии

Предположения множественной регрессии отличаются от предположений одномерной регресии, по сути, только в формулировках.

1. Зависимость между переменной Y и вектором независимых переменных (X_1, X_2 .. X_n) носит линейный характер.
2. Вектор независмых переменных (X_1, X_2 .. X_n) не является случайным. Кроме того, не существует линейной связи между двумя или более независимыми переменными.
3. Математическое ожидание остатка по каждой переменной равно 0.
4. Вариация остатков постоянна для всех наблюдений.
5. Остатки не имеют автокорреляции и являются нормально распределенными.

1.9. Мультиколлинеарность

Что происходит, когда вы пытаетесь делать регрессию с мультиколлинеарными независимыми переменными? Постарайтесь собрать вместе все сохранившиеся в памяти сведения из линейной алгебры (отговорка о том, что вы не изучали в ВУЗе высшую математику, не принимается — линейная алгебра, хоть и в минимальном объеме, но входит в курс средней школы).

Представьте, что вы пытаетесь разложить вектор по базису из двух других векторов; проблема же состоит в том, что вы не уверены ни в модуле (длине), ни в направлении векторов. У вас есть такие сведения: примерное направление (плюс-минус 15 градусов) и примерный модуль (1 плюс-минус 0.1). Теперь если два эти вектора (их «средние» направления) направлены перпендикулярно друг другу, то коэффициенты, полученные при разложении вектора по базису из двух таких векторов, не будут отличаться существенным образом.

С другой стороны, представив, что угол между «средними» направлениями векторов составляет 15 градусов, мы получим, что реальный угол между векторами может находиться в пределах от -15 до +45 градусов. Понятно, что пытаясь разложить наш вектор — да и вообще любой вектор — по такому, с позволения сказать, базису, ничего хорошего мы не получим.

Мультиколлинеарность не обязательно проявляется в случае наличия корреляции между двумя или несколькими независимыми переменными; однако, наличие этой корреляции может подсказать о возможности существования такой неприятной проблемы, как мультиколлинеарность.

Кстати, а чем, собственно, она плоха, мультиколлинеарность? В общем, наше описание «на пальцах» достаточно хорошо отражает главную проблему: снижение точности определения коэффициентов.

1.10. «Раскапывание» данных и игра в R^2

Выше уже приведен пример «раскапывания» данных (data mining). Пример, однако, не описывает, что нужно делать, чтобы его избежать.

Думаю, рассказ о том, какие общие принципы можно применять при построении моделей, избавит вас от многочисленных ошибок.

1. Прежде всего, ваша модель должна быть основана на выборе разумных и логичных переменных. Никаких «цен на раков», никаких «пятен на Солнце» — только те переменные, влияние которых на результат согласуется со здравым смыслом.
2. Выбранная функциональная форма должна соответствовать существующей природе зависимости между переменными. Например, если вы строите зависимость стоимости актива от времени (еще раз — это не обязательно хорошая идея, строить регрессию чего бы то ни было на время), правильнее применять логарифмическое преобразование.
3. Модель должна быть скупой (англоязычные специалисты говорят parsimonious). Чем меньше переменных — тем лучше (по крайней мере, тем большую часть вариации объясняет каждая из них).
4. Следует проверить модель на удовлетворение всех предположений линейной регрессии.
5. При построении модели, если вам приходится перебирать много переменных, используйте процедуры коррекции (скорректированный $\mathbf{R^2}$, коррекция Бонферрони, и т.п.)
6. Модель должна быть проверена на данных, отличных от тех, на базе которых она была создана; при этом она должна показать хорошие результаты.

При соблюдении этих условий неважно, будете вы строить модель «снизу вверх» или «сверху вниз» (то есть будете ли вы начинать с одной-двух переменных, постепенно увеличивая их число, или с максимально большого числа переменных, постепенно избавляясь от лишних). Что касается персональных предпочтений, я стою за построение модели «снизу вверх» — хотя, чего греха таить, модель «сверху вниз» иногда привлекательней с точки зрения экономии времени.

1.11. Бинарные переменные (dummy variables)

Бинарными в статистике называют переменные, которые могут принимать только 2 значения — 0 или 1. Именно такими переменными описывают, например, сезон, месяц года, день недели, наличие страховки и т.д.

Пример использования бинарных переменных приведен на листе «Магазин» (см. Приложение 3) книги «Примеры». В этом случае были использованы переменные «Май», «Декабрь» и «Бюджет». Очевидно, первые две из них принимают значение 1 в

соответствующем месяце, а переменная «Бюджет» принимает значение 1 только в случае, если — как нам стало известно пост-фактум — были перечислены деньги бюджетным организациям.

При этом понятно, что мы могли бы ввести 11 переменных — по числу «линейно-независимых» месяцев (чтобы лучше это понять, попробуйте создать 12 переменных по числу месяцев — вероятнее всего, Excel присвоит одной из них коэффициент 0).

Более того, введя, скажем, 46 переменных, мы сможем описать 47 наблюдений абсолютно точно. Разумеется, такая модель не будет иметь никакой предсказательной силы.

1.12. Использование Excel для анализа линейной регрессии на нескольких переменных

По большому счету, использование Excel для анализа регрессии на несколько переменных очень мало отличается от случая одной переменной. Формат функции тот же: ЛИНЕЙН (Известные_Y, Известные_X, A, Статистика). Единственное отличие состоит в том, что в данном случае Известные_X — несколько столбцов (обязательно соседних).

Напомним, что мы будем пользоваться этой функцией в формате ЛИНЕЙН (Известные_Y, Известные_X, 1,1).

Кроме того, напомним, что для получения всех результатов регрессии, нам необходимо следующее.

1. Создать в ячейке указанную выше формулу.
2. Пометить прямоугольник высотой 5 строк и шириной k+1 столбцов (k — количество независимых переменных).
3. Нажать сначала клавишу F2, а затем — одновременно — Ctrl+Shift+Enter.
4. Если вы все сделали правильно, появятся результаты регрессии Y на X.

Некоторое неудобство заключается в том, что Excel выводит коэффициенты в порядке, обратном тому, в котором они находились. Поясним (приведены исходные условия и результаты двумерной регрессии из примера с количеством КПК на 1000 человек населения).

Исходные данные:

Возраст (середина)	Заработная плата	КПК на 1000 человек
17,5	19	57
22,5	32	78
27,5	40	99
32,5	50	121

37,5	55	147
42,5	65	168
47,5	67	175
52,5	70	184
57,5	72	164
62,5	70	145

Результаты регрессии:

З/плата	Возраст	Интерсепт
3,645733	-1,77145	7,7884266
0,556114	0,676043	10,213112
0,960088	9,841037	#N/A
84,19386	7	#N/A
16307,68	677,922	#N/A

Раздел 2. Линейное программирование

2.1. Цели использования

Линейное программирование используется для оптимизации решения (детерминистских — то есть имеющих точное решение) задач.

За этой формулировкой скрывается удобный инструмент, с помощью которого компании решают целый ряд задач.

Какие же задачи чаще всего решаются с его помощью?

Определение «product mix» — набора выпускаемой продукции. Большинство производственных компаний могут производить целый набор продуктов. Однако каждый продукт требует различного количества исходных материалов и трудовых затрат. Разумеется, прибыль, генерируемая каждым продуктом, также отличается. Соответственно, менеджер такой компании должен принять решение о том, какие продукты производить, — с целью максимизировать прибыль или удовлетворить возможный спрос при минимальных расходах.

Производство. Например, в некоторых специальных печатных платах приходится проделывать сотни или тысячи отверстий, чтобы разместить все компоненты, которые должны быть на ней размещены. Чтобы изготовить такие платы, сверлильный станок с программным управлением нужно запрограммировать на сверление в первом заданном месте, затем передвинуть сверло к месту следующего отверстия и просверлить его. Так как отверстий много, процесс повторяется сотни и тысячи раз; поэтому производители

печатных плат существенно выиграют от определения порядка сверления, который минимизирует время позиционирования инструмента (вообще говоря, необходимо минимизировать общее время обработки; но мы считаем, что на скорость сверления мы повлиять не можем).

Финансовое планирование. Например, вы хотите оптимизировать ваши вложения, перенося результаты, полученные в прошлом, на будущее. При этом возникает следующая задача: поскольку уровень вашего потенциального дохода связан с уровнем риска, вы, задавшись определенным уровнем риска, хотели бы получить максимально возможный при данном уровне риска доход.

2.2. Постановка задачи

Математическая постановка задачи выглядит следующим образом:

Найти max $\{F (X_1,X_2,X_3,..X_n)\}$ при условиях:

$$f_1 (X_1,X_2,X_3,..X_n) \geq 0$$
$$f_2 (X_1,X_2,X_3,..X_n) \geq 0$$
$$f_3 (X_1,X_2,X_3,..X_n) \geq 0$$
$$.$$
$$.$$
$$.$$
$$f_n (X_1,X_2,X_3,..X_n) \geq 0,$$

где $X_1 \ldots X_n$ — некие переменные (например, набор может быть таким: количество продукции по видам, расход материалов, трудовые затраты; то есть наш «ответ», вообще говоря, входит в число переменных), а F, $f_1 \ldots f_n$ — функции от этих переменных.

При этом функция F носит название целевой функции (например: прибыль, пройденное расстояние, ожидаемый доход и т.д.).

Условие $f_k (X_1,X_2,X_3,..X_n) \geq 0$ иногда называют ограничением; жесткие условия вида $f_k (X_1,X_2,X_3,..X_n) = 0$ мы рассматривать не будем — главным образом, ввиду того, что данные условия плохо воспринимаются надстройкой Excel «Поиск решения» (мы, однако, рассмотрим способ заставить Excel поверить в то, что он работает с «мягким» условием, заменяющим жесткое для наших целей).

Понятно, что в такой постановке общность задачи не ограничена: так, например, для того чтобы найти min $\{F (X_1,X_2,X_3,..X_n)\}$, нужно искать max $\{- F (X_1,X_2,X_3,..X_n)\}$.

Аналогичным образом мы можем поступить и с другими функциями, производя алгебраические действия, не изменяющие равенств.

Пример постановки задачи

Теперь сформулируем типичную **задачу**.

Компания «Русские джакузи» производит и продает (разумеется, производство не может быть самоцелью — во всяком случае, в капиталистической экономике) два вида ванн-джакузи: Аква-Спа и Гидро-Люкс. Руководитель компании должен принять решение в отношении количества ванн, которые нужно произвести в следующем квартале.

При этом производство устроено следующим образом: компания покупает отформованные пластиковые ванны у местного поставщика и добавляет насос, трубы и форсунки, чтобы превратить обычную ванну в джакузи. Поставщик может продать любое, сколь угодно большое количество ванн.

Для экономии издержек, в обоих типах ванн используется один и тот же тип насоса. В следующем квартале поставщик насосов может поставить только 200 штук насосов требуемого типа.

С точки зрения производства, главное отличие между джакузи заключается в количестве рабочего времени и длине используемых труб.

В таблице приведены данные для Аква-Спа и Гидро-Люкс:

	Аква-Спа	**Гидро-Люкс**
Длина труб	12 метров	16 метров
Затраты рабочего времени	9 часов	6 часов
Операционная прибыль	350 долларов	300 долларов

Количество рабочего времени достаточно жестко регулируется профсоюзом, поэтому максимальное число рабочих часов составит не более 1566. Еще одно ограничение накладывается поставщиком труб: он не может поставить более 2000 метров (плюс 880 метров есть на складе; в дальнейшем поставки возобновятся).

Вся выпускаемая продукция находит своего покупателя. Менеджер уверен, что это продлится и в будущем.

Вопрос: какое количество ванн Аква-Спа и Гидро-Люкс нужно произвести, чтобы максимизировать прибыль за период?

(Обратите внимание: прямые затраты уже учтены в операционной прибыли, то есть нам нет нужды знать стоимость рабочего часа и метра трубы).

Распишем порядок решения задачи подробно.

1. *Понимаем задачу.* В данном случае, задача достаточно очевидна. В большинстве ситуаций, с которыми вам придется столкнуться, само формулирование задачи —

во-первых, процесс непростой, а во-вторых — уже половина решения; что уж говорить о понимании.

2. *Определяем переменные в решении.* В нашем случае, такими переменными будут X_1 и X_2 — количество Аква-Спа и Гидро-Люкс соответственно.
3. *Определяем целевую функцию.* Поскольку мы назвали наши переменные X_1 и X_2, целевая функция будет выглядеть следующим образом:
    i. $F(X_1,X_2) = 350X1 + 300X2$, а задача — найти $\max \{ F(X_1,X_2) = 350X_1 + 300X_2\}$.
4. *Определяем налагаемые ограничения.*
    i. Ограничение по количеству насосов будет выглядеть следующим образом:
    $X_1 + X_2 \leq 200$
    Ограничение по количеству рабочих часов:
    $9X_1 + 6X_2 \leq 1566$
    Ограничение по длине трубы:
    $12X_1 + 16X_2 \leq 2880$
5. *Определяем (возможные) ограничения на решение.* В частности, очевидно, что количество произведенных ванн должно быть положительно (для пытливых умов: отрицательное количество произведенных ванн значило бы, что вы покупаете готовые джакузи на рынке и разбираете их на составляющие с тем, чтобы использовать трубы и насос для производства продукции с наибольшей маржой).
    i. $X_1 \geq 0$
    $X_2 \geq 0$
    ii. Такие условия обычно называют условиями «неотрицательности».

Решение задач линейного программирования в Excel

Как выглядит решение нашей задачи **на графике**?

Каждое ограничение представляет собой прямую, отсекающую часть плоскости. Таким образом, область возможных решений ограничена снизу и слева осями OX и OY, а сверху — ломаной линией, образованной, в свою очередь, линиями, проведенными в соответствии с нашими ограничениями.

См. рис. 7.

Итак, мы (графически) определили область возможных решений. Как поступить дальше? Если наша функция линейна (то есть содержит переменные только в первой степени — в частности, не содержит произведений переменных между собой), не сложно догадаться, что максимума она может достигать либо в одной из вершин, образованных пересечением линий ограничений, либо на одном из отрезков, образованных этими

вершинами (и тогда в любой точке отрезка достигается максимум). Таким образом, в нашем случае мы можем, не мудрствуя лукаво, перебрать все 5 вершин с тем, чтобы найти оптимальное решение.

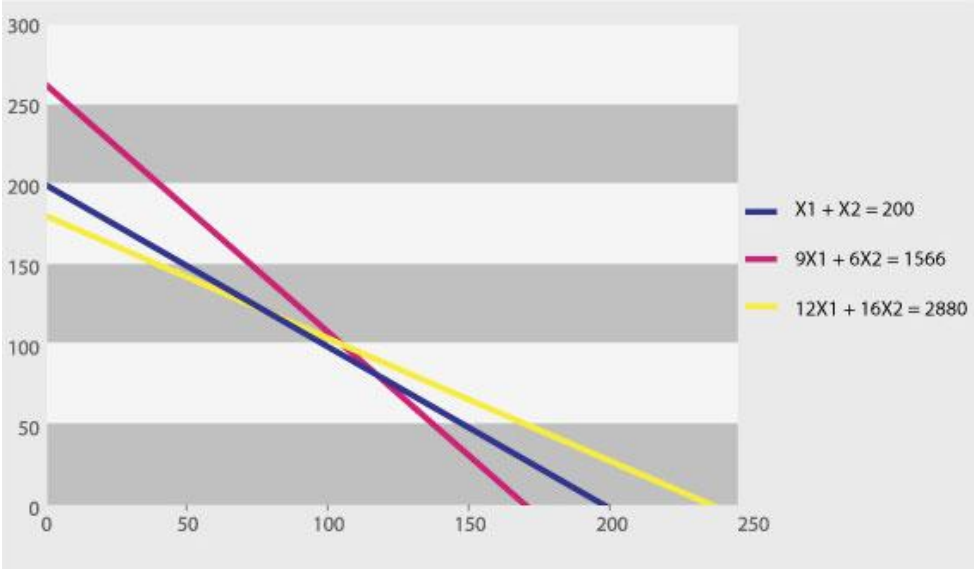

Рис. 7

Однако далеко не всегда мы сможем обойтись всего двумя переменными; здесь возникает та же проблема, что и во множественной регрессии, — нам все сложнее и сложнее будет представить область возможных решений; и даже поиск точек пересечения — если мы решим отказаться от графического построения таких областей — будет становиться все более непростой задачей.

К счастью, нам на помощь, как всегда, придет Excel — с его надстройкой (add-in) «Поиск решения» (Solver — в англоязычном варианте).

Комментарий: Надстройка «Поиск решения», хотя и входит в стандартный инсталляционный комплект, не устанавливается автоматически; для установки надстройки необходимо совершить следующие действия.

1. Выбрать пункт меню «Сервис» (Tools).
2. Выбрать подпункт «Надстройки» (Add-ins).
3. В появившемся списке отметить нужные надстройки (в ваших интересах установить, как минимум, «Поиск решения» — Solver).

Ниже приведена таблица, в которой организованы наши данные.

	Аква-Спа	Гидро-Люкс	Итого
Количество	0	0	0
Маржа	350	300	0
Трубопровод	12	16	0
Рабочее время	9	6	0

Чтобы сократить таблицу, для подсчета в колонке «Итого» — во всех строках, кроме строки количество, где использовалась обычная сумма — мы воспользовались функцией СУММПРОИЗВ (в англоязычном варианте, соответственно, SUMPRODUCT). Эта функция вычисляет сумму попарных произведений двух векторов равной длины (скалярное произведение, если угодно). То есть, например, СУММПРОИЗВ(A1:A2;B1:B2) = A1×B1 + A2×B2. Таким образом, в строке «Маржа» в колонке «Итого» стоит сумма попарных произведений количества и маржи — а это и есть операционная прибыль, наша целевая функция.

Теперь можно вызвать «Поиск решения», выбрав в меню «Сервис» пункт «Поиск решения». В появившемся окне диалога необходимо сделать следующее.

1. Указать целевую ячейку (в нашем случае — итоговая маржа).
2. В подстрочке «Равной» выбрать необходимый пункт (в нашем случае, разумеется, максимум).
3. В окошке «Изменяя ячейки» выбрать те ячейки, которые мы собираемся изменять (в нашем случае — количество джакузи разных видов).
4. В окошке «При условии» нажать «Добавить» и по одному ввести наши условия (объяснение см. выше):

$X_1 + X_2 \le 200$
$9X_1 + 6X_2 \le 1566$
$12X_1 + 16X_2 \le 2880$

Кроме того, не стоит забывать условия неотрицательности ($X_1 \ge 0$ и $X_2 \ge 0$) — мы же не хотим разбирать готовые джакузи на составные части.

Закончив с ограничениями, нажимаем кнопку запуска — и через несколько секунд получаем готовый ответ.

	Аква-Спа	**Гидро-Люкс**	**Итого**
Количество	122	78	200
Маржа	350	300	66100
Трубопровод	12	16	2712
Рабочее время	9	6	1566

Легко убедиться, что все ограничения выполнены.

Виды ограничений и рекомендации

Виды ограничений и особые состояния

Избыточные ограничения. Например, при наличии двух ограничений: $X_1 + X_2 \le 200$ и $X_1 + X_2 \le 300$, последнее явно избыточно.

Неограниченное решение. Например, если единственное существующее ограничение $X_1 \leq 200$, мы сможем сколь угодно далеко продвинуться вдоль оси X_2.

Невозможные (несовместные) условия. Например, $X_1 + X_2 \leq 200$ и $X_1 + X_2 \geq 500$. Очевидно, что возможные области лежат по разные стороны от полосы, разделяющей плоскость X_1X_2.

Альтернативные решения. Как уже отмечалось, в случае, если оптимальное решение приходится не на вершину, а на ребро (а в многомерном случае — на часть плоскости или гиперплоскости), то мы вправе выбирать из множества решений. В этом случае разумно установить дополнительный критерий.

Рекомендации

1. «Поиск решения» — отнюдь не идеальное средство. Если нет возможности приобрести альтернативы, относитесь к результатам с осторожностью, если не сказать — с предубеждением.
2. Не используйте жесткие равенства. Вместо $X_1=10$ лучше использовать, например, такое ограничение: $(X_1 - 10)^2 \leq 0.5$. Затем это ограничение можно (постепенно!) ужесточать.
3. Заметьте: в предыдущем случае был использован *квадрат*, а не *модуль* (абсолютная величина). Решая нелинейные задачи, вы сэкономите на этом массу времени — и себе, и компьютеру.
4. С особой осторожностью следует использовать ограничения типа ЦЕЛОЕ и ДВОИЧНОЕ. Эти ограничения существенно повышают порядок задачи — и, соответственно, увеличивают время расчета.
5. Если «Поиск решения» сообщает о том, что решение не найдено, первым делом проверьте ваши ограничения. Затем попробуйте поменять стартовые величины.
6. Даже если «Поиск решения» сообщает о том, что решение найдено, попробуйте «сдвинуть» решение. Шанс, что таким образом вы найдете лучшее решение, довольно велик.
7. Чтобы окончательно убедиться, что решение оптимально, можно воспользоваться методом Монте-Карло. Его-то мы и будем изучать в следующем разделе.

Раздел 3. Метод Монте-Карло

3.1. Стохастические задачи (детерминистские)

Решение задачи методом линейного программирования — да и вообще, любым детерминистским методом — обладает одним существенным недостатком: оно не учитывает возможные *варианты* развития событий.

Представьте себе, что ваша компания (та самая «Русские Джакузи» из предыдущей главы) должна ежемесячно погашать кредит в размере 1 600 000 рублей. Компания рассчитала оптимальное количество выпускаемой продукции (напоминаю, операционная прибыль в этом случае равна 66 100 долларов). Поскольку текущий курс доллара составляет 25 рублей за доллар, руководство компании считает, что запаса финансовой устойчивости вполне достаточно (в конце концов, 1652500 — 1600000 = 52500 — запас хоть и небольшой, но пока еще положительный).

Теперь представьте, что в результате кратковременного изменения спроса компании не удалось продать часть произведенной продукции, что привело к уменьшению маржи на 1600 долларов. Но это не все; поскольку курс доллара снизился (до 24.25 рублей за доллар), это еще больше ухудшило результат.

Соответственно, в этом месяце баланс наличности компании равен:

(66100-1600) ×24.25 - 1600000 = 1564125 - 1600000 = - 35875 рублей.

Можно рассчитывать, что такой отток денежных средств не вызовет банкротства компании; однако несколько месяцев таких результатов — и его перспектива будет вполне зримой.

Согласитесь, проделав все упражнения с линейным программированием, вам было бы обидно увидеть свое детище банкротом — только из-за того, что вы не учли возможность одновременного снижения спроса на 2-3 процента и снижения обменного курса на те же самые 3 процента?

Каким образом мы можем **модифицировать детерминистские методы** с тем, чтобы избавить их от этого недостатка?

3.2. Модифицированные детерминистские и стохастические методы

Как обычно, существует несколько способов:

- best/worst case анализ (анализ лучшего/худшего сценария);
- анализ сценариев;
- симуляция методом Монте-Карло.

Каждому из этих методов присущи свои достоинства и недостатки.

Так, **анализ лучшего/худшего сценариев** не учитывает вероятности этих сценариев.

Представьте, что вы покупаете лотерейный билет. Худший сценарий — отсутствие выигрыша, лучший — выигрыш в 1 миллион долларов. Если вы не знаете связанных с этим вероятностей, выглядит заманчиво, не так ли?

Анализ сценариев (здесь имеется в виду перебор большого числа комбинаций факторов) требует больших затрат рабочего времени, и, по большому счету, не избавляет от проблемы с незнанием вероятностей исхода.

Представьте, что вы делаете модель работы нефтяной компании. Новое бурение (а вы обязаны его осуществлять как по условиям лицензии, так и для поддержания уровня добычи) может дать результаты от 0 (сухая скважина) до, например, 1000 тонн в сутки. Таких новых скважин — 70 за год; это уже 70 параметров. Плюс цена на нефть, на каждый вид нефтепродуктов, курс валюты, экспортная пошлина, спрос, строительство новых заправок, мини-заводов — всего, допустим, наберется 200 параметров. Если вы захотите придать каждому из них 5 уровней значения, полное число вариантов будет 5200. Это... много. Даже если вы уменьшите число параметров до 10, а число уровней значения — до 2, то все равно перебрать 1024 (2^{10}) вариантов вручную — задача почти безнадежная.

Метод Монте-Карло, в свою очередь, избавит нас от необходимости определять вероятности определенных исходов — он это сделает за нас.

Не бывает методов без недостатков. Метод Монте-Карло тоже их не лишен: он требует большой вычислительной мощности и, по сути, не позволяет решать задачи с большим числом шагов. Если моделируя 1 период (периодом может быть месяц, квартал, год и т.д.), мы должны сделать, например, 1000 симуляций, то два периода потребуют 1000^2, три — 1000^3 и т.д.

3.3. Постановка задачи

В духе самого метода, не будем стремиться к математической точности постановки задачи; вместо этого, опишем этапы формирования модели.

1. Как обычно, прежде всего, нужно понять, что мы, собственно, моделируем.
2. Определить переменные, которые мы будем использовать для симуляции.
3. Создать детерминистскую модель, результаты работы которой будут верными для неизменного набора входящих переменных.
4. Определить, какой тип распределения следует использовать, моделируя изменение данных переменных (например, если мы хотим моделировать спрос, нам следует использовать распределение Пуассона, если средняя величина спроса невелика).
5. (Этот шаг можно пропустить) Создать или найти подходящий генератор (псевдо)случайных чисел. Обратите внимание: если вы будете модифицировать свою модель, то весьма желательно, чтобы генератор случайных чисел обладал свойством *воспроизводимости* (то есть при одном и том же стартовом числе он должен выдавать одну и ту же последовательность «случайных» чисел — для того, чтобы вы были уверены, что изменение результата вызвано исключительно изменением модели, а не изменением набора входящих данных).

6. Поскольку, как правило, генераторы случайных чисел выдают псевдослучайное число, равномерно распределенное на промежутке от 0 до 1, нужно построить распределение, обратное тому, в соответствии с которым распределены наши случайные величины (в ходе реализации модели в Excel этот шаг будет понятнее).
7. Обеспечить регистрацию результатов каждого прогона.
8. Задаться определенным числом симуляций и запустить модель.
9. Увеличить число симуляций, принимая во внимание, что точность модели увеличивается пропорционально квадратному корню из числа симуляций.
10. Когда точность решения станет достаточной, обработать результаты — и, наконец, принять решение.

3.4. Реализация решения в Excel

1. Мы будем рассматривать описанный выше случай — «Русские Джакузи»; при этом нашей целью будет установить вероятность отрицательного денежного потока в текущем периоде.

Напомним данные:

	Аква-Спа	**Гидро-Люкс**
Длина труб	12 метров	16 метров
Затраты рабочего времени	9 часов	6 часов
Операционная прибыль	350 долларов	300 долларов

Кроме того, будем считать, что курс доллара по отношению к рублю распределен нормальным образом с параметрами (25, 0.5) — то есть, среднее значение равно 25 рублям за доллар, а стандартное отклонение равно 0.5. Далее, спрос на Аква-Спа и Гидро-Люкс тоже распределен нормально с параметрами (122, 2) и (78, 2) соответственно.

2. Понятно, что для симуляции мы будем использовать 3 переменных — курс доллара и спрос на ванны обоих типов.
3. Детерминистская модель крайне проста; ее создание описывать не будем (см. Приложение 4).
4. Поскольку тип распределения мы приняли заранее, этот шаг можно пропустить.
5. К счастью, создатели Excel позаботились о генераторе случайных чисел (правда, не без недостатков). В русском варианте эта функция называется СЛЧИС() (в английском, соответственно, RAND()).
6. Для генерации нормально распределенного случайного числа следует воспользоваться функцией НОРМОБР (вероятность; среднее; стандартное отклонение) (в английском варианте — NORMINV (…)).
7. и 8. В данном случае при создании модели мы не пользовались макросами; чтобы получить 1000 различных исходов, нам достаточно просто скопировать первую

строчку требуемое количество раз (разумеется, можно применить автозаполнение). Теперь, чтобы произвести еще одну симуляцию 1000 прогонов, нам достаточно просто нажать F9 (настоятельно рекомендую отключить автоматический пересчет таблиц!!).

9. Нажимая F9, вы, наблюдая за результатами (самым важным для нас является процент исходов с отрицательным денежным потоком), можете принять решение о необходимости увеличения количества прогонов.
10. Оценив количество исходов с отрицательным денежным потоком, вы, возможно, примете решение об увеличении акционерного капитала.

Рекомендации

1. Если ваша таблица велика, отключайте автоматический пересчет — это сбережет вам время и нервы.
2. Помните, что вы можете самостоятельно создать любое распределение случайной величины.
3. Если вам приходится моделировать коррелирующие случайные величины, следует применять специальные методы (например, факторизацию по Холецкому).
4. В случае таблиц, требующих объемного вывода (то есть вывода многих параметров), проще использовать макросы.
5. Экспериментируйте! Чем больше вы будете развлекаться с генераторами Монте-Карло, тем лучше вы почувствуете метод.

3.5. Метод Монте-Карло: модифицированный способ с использованием макроса

К сожалению, иногда ваша таблица будет настолько сложна, и будет иметь такое количество входящих — моделируемых — переменных, что вам придется использовать несколько иной способ, регистрируя для каждого варианта набор входящих переменных и набор исходящих переменных (то есть тех, на которых потом будет строиться ответ).

Помните задачу о мальчике, продающем газеты? Настало время ее решить.

Напомним вкратце условия.

Вы продаете популярную газету (Financial Times, например). Вы покупаете газету за 1 доллар, а продаете за 2. При этом редакция газеты готова забрать все экземпляры, оставшиеся непроданными на следующий день, по 50 центов. Вы подсчитали, что средний спрос на газету составляет 100 штук; при этом стандартное отклонение спроса составляет 20 штук. Какой заказ следует делать, чтобы максимизировать прибыль?

Чтобы проверить, какой заказ будет оптимальным, поступим следующим образом: будем генерировать последовательно 1000 «дней» спроса на нашу газету. При этом чтобы быть уверенными в том, что разница в прибыли вызвана исключительно разницей в величине заказа, мы будем использовать один уровень спроса для всех продавцов.

Как было описано выше, мы будем регистрировать каждый прогон (вы увидите, как это сделано, запустив макрос. Кстати, написание макросов — это несложный процесс, который может быть освоен без знания навыков программирования).

Модель реализована в файле News Vendor Model.xls (Приложение 5) (именно так называется эта модель в теории управления операциями). Эта задача известна очень широко, а выводы из нее и ее усложненных модификаций распространены на многие сферы деятельности, где есть проблема продажи товара или услуги, имеющей весьма ограниченный «срок годности», — как в случае с газетой, авиабилетом или номером в гостинице.

Для экономии места в пересылаемом файле нет результатов симуляции; в моем (единственном) запуске наилучшего результата можно было добиться при уровне заказа 109.

Здесь, очевидно, нужно сделать оговорку: разумеется, данную задачу можно было решить так же, как и предыдущую, — без использования макросов. Способ с использованием макроса — ничем не отличающийся математически — еще один вариант реализации модели, не меняющий принципиально ее характеристик.

Заключение

Надеюсь, вам удалось «пробраться» через довольно сложный курс. Как и говорилось в самом начале, курс совершенно не претендует на полноту; главной задачей было дать вам направления для поиска и постараться привить вкус к численному решению бизнес-задач, будь то детерминистские или стохастические способы.

Уверен, что рассмотренные в курсе методы помогут вам в вашей ежедневной работе; не забывайте, что каким бы бизнесом вы ни занимались, результаты работы этого бизнеса могут быть улучшены и/или стабилизированы с помощью этих методов.

Напоследок позвольте пожелать вам удачи и всяческих успехов — в том числе, и в бизнесе!

Глоссарий

А

Автокорреляция

корреляция между значениями временно́го ряда и значениями с предыдущего шага (т.е., между X_t и X_{t-1}). Если такая корреляция больше нуля, то, если значение нашего временно́го ряда на определенном шаге больше (меньше) нуля, то значение на следующем шаге с большей вероятностью тоже больше (меньше) нуля. Соответственно, если автокорреляция отрицательна, то положительное значение на одном шаге с некоторой вероятностью означает отрицательное на следующем шаге и наоборот.

И

Интерсепт (пересечение, точка пересечения)

значение зависимой переменной в точке, где все независимые переменные принимают значение 0.

М

Мультиколлинеарные переменные

нарушение предположений линейной регрессии, когда некоторые из переменных или их линейные комбинации коррелируют между собой (частично, но не на 100%).

О

Остаток

разность между реально наблюдавшейся величиной и значением, предсказанным уравнением регрессии.

С

Стандартная ошибка

мера возможной ошибки оценки. Например, если стандартная ошибка определения наклона S_b = 1, то мы можем сказать (если оценка наклона равна 10), что истинный наклон с вероятностью 95% попадает в промежуток 10±2 (т.е., находится в промежутке между 8 и 12). Соответственно, если **стандартная ошибка прогноза** S_e = 2, то мы можем ожидать, что (при условии соблюдения всех предположений линейной регрессии, в частности, отсутствия гетероскедастичности), истинные значения будут находиться в полосе ±4 относительно линии регрессии.

У

Уровень значимости α

заранее задаваемая вероятность совершения ошибки 1-го типа (т.е., отклонение нулевой гипотезы, когда в действительности она верна).

R

R^2

показатель «качества» регрессии. В случае линейной регрессии, совпадает с квадратом коэффициента корреляции. В случае множественной регрессии - $\mathbf{R^2}$ представляет собой процент вариации зависимой величины, объясняемый регрессией.

Список рекомендуемой литературы

1. Джекел Питер. Применение методов Монте-Карло в финансах (+ CD-ROM) (Monte-Carlo Methods in Finance). — Издательство: Интернет-трейдинг, 2004
2. Минько А. А. Статистика в бизнесе. Руководство менеджера и финансиста. — Издательство: Эксмо, 2008
3. Норман Р. Дрейпер, Гарри Смит. Прикладной регрессионный анализ (Applied Regression Analysis). — Издательства: Вильямс, Диалектика, 2007
4. Сигел Эндрю. Практическая бизнес-статистика. (Practical Business Statistics). — Издательство: Вильямс, 2007 г, ISBN 5-8459-0306-8.
5. Соколов Г. А., Гладких И. М. Математическая статистика. — Издательство: Экзамен, 2004