Майкл Х. Херцог, Грегори Френсис, Аарон Кларк

Статистика и планирование эксперимента

для непосвященных

ДМК ПРЕСС
издательство

Непонимание статистики – важная проблема в нашем обществе. Благодаря компьютерным технологиям собирать статистические данные стало проще, но главную задачу – правильно обработать результаты – по-прежнему берет на себя человек. Из этой книги вы узнаете, как использовать и интерпретировать статистику и статистические данные в различном окружении.

Среди рассматриваемых тем:

- основные понятия и принципы статистики;
- наиболее распространенные статистические критерии;
- множественная проверка гипотез;
- планирование эксперимента;
- метастатистика (статистическое исследование статистики).

Издание пригодится тем, кто хочет понять принципы статистики и научиться интерпретировать ее результаты, не вдаваясь в математические детали вычислений. Для изучения материала требуется минимальный уровень математической подготовки.

ДМК ПРЕСС
ИЗДАТЕЛЬСТВО
www.дмк.рф

Интернет-магазин: www.dmkpress.com

Оптовая продажа: КТК "Галактика" books@alians-kniga.ru

ISBN 978-5-93700-195-5

Майкл Х. Херцог, Грегори Фрэнсис, Аарон Кларк

Статистика и планирование эксперимента для непосвященных

Как отучить статистику лгать

Michael H. Herzog • Gregory Francis • Aaron Clarke

Understanding Statistics and Experimental Design

How to Not Lie with Statistics

Springer

OPEN

Майкл Х. Херцог • Грегори Фрэнсис • Аарон Кларк

Статистика и планирование эксперимента для непосвященных

Как отучить статистику лгать

ДМК Пресс
издательство

Москва, 2023

УДК 519.242
ББК 22.183
Х39

Майкл Х. Херцог, Грегори Фрэнсис, Аарон Кларк
Х39 Статистика и планирование эксперимента для непосвященных: Как отучить статистику лгать / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2023. – 174 с.: ил.

ISBN 978-5-93700-195-5

Непонимание статистики – важная проблема в нашем обществе. Благодаря компьютерным технологиям собирать статистические данные стало проще, но главную задачу – правильно обработать результаты – по-прежнему берет на себя человек. Из этой книги вы узнаете, как использовать и интерпретировать статистику и статистические данные в различном окружении. Рассмотрены основные понятия и принципы статистики, наиболее распространенные статистические критерии, множественная проверка гипотез, планирование эксперимента, а также метастатистика.

Издание пригодится тем, кто хочет понять принципы статистики и научиться интерпретировать ее результаты, не вдаваясь в математические детали вычислений. Для изучения материала требуется минимальный уровень математической подготовки.

УДК 519.242
ББК 22.183

ISBN (анг.) 978-3-030-03498-6
ISBN (рус.) 978-5-93700-195-5

Оглавление

Предисловие от издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте *www.dmkpress.com*, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу *dmkpress@gmail.com*; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу *http://dmkpress.com/authors/publish_book/* или напишите в издательство по адресу *dmkpress@gmail.com*.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу *dmkpress@gmail.com*, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу *dmkpress@gmail.com*.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Предисловие

Наука, общество и статистика

Современный мир до краев заполнен статистикой. Статистика знает, что мы едим, как тренируемся, с кем дружим, как учим своих детей и какие принимаем лекарства. Очевидно, что статистика вездесуща, как и – к глубокому сожалению – окружающие ее домыслы. В главе 1 мы расскажем о том, как судьи выносят неправильные приговоры, – отправлять человека в тюрьму или нет, – поскольку не понимают даже основ статистики. Мы покажем, что пациенты совершают самоубийство, потому что врачи не умеют интерпретировать результаты анализов. Ученые зачастую ничуть не лучше. Мы знаем коллег, которые слепо доверяли результатам статистических программ, даже когда они не имели ни малейшего смысла. Встречались нам и опубликованные научные статьи, в которых результаты не согласуются с теоретическими выводами авторов.

Есть старое изречение (иногда его приписывают Марку Твену, но, по-видимому, оно все же старше): «Существует три вида лжи: ложь, наглая ложь и статистика». Мы вынуждены признать, что в нем имеется зерно истины. Люди часто неправильно пользуются статистическим анализом. Быть может, до прямой лжи (высказывания заведомо ложного утверждения) и не доходит, но статистика зачастую только запутывает, а не проясняет дело. Название этой книги отражает наши усилия повысить качество использования статистики так, чтобы те, кто занимается анализом, научились лучше интерпретировать свои данные, а те, кто читает статистические результаты, лучше их понимали. Понимание ключевых идей статистики поможет осознать, что многие научные результаты по сути своей бессмысленны, и объяснит, почему многие эмпирические науки в настоящее время сталкиваются с кризисом воспроизводимости. *Вычисление* статистики может оказаться очень трудным делом (отсюда и сложные программы, и толстые книги с глубокими теоремами), но ясное понимание основных принципов статистики вполне доступно каждому.

В 2013 году, вдохновленные этими идеями, мы начали читать курс в Федеральной политехнической школе Лозанны, Швейцария, посвященный концептуальным основам статистики и планирования эксперимента. С годами курс стал довольно популярным, его стали

посещать студенты, обучающиеся биологии, нейронаукам, медицине, генетике, психологии и биоинженерии. Обычно такие студенты уже прослушали один или несколько курсов по статистике, на которых им преподавали *детали* статистического анализа. На нашем же курсе и в этой книге акцент сделан на *базовых* принципах этого анализа; мы хотим кратко объяснить, что это такое, и привить читателю понимание возможностей и ограничений анализа.

Об этой книге

Предварительные знания и цель. Как уже было сказано, непонимание статистики стало важной проблемой в нашем обществе. И одно из ее проявлений в том, что вычислить статистические показатели теперь так просто, что наличие хорошего образования кажется ненужным. Однако же все в точности наоборот. Простота использования статистических программ позволят выполнять анализ данных, не понимая, что программа делает и как интерпретировать ее результаты. Итог – несостоятельные выводы. Читатель, вероятно, удивится тому, насколько проблема серьезна и сколь велико количество бесполезных исследований. И еще, наверное, с удивлением узнает, что даже такие базовые термины, как *p*-значение, означают совсем не то, что многим кажется.

Главная цель этой книги – кратко и по существу изложить основы математической статистики. Понимание этих основ подготовит читателя к правильному восприятию и критической оценке научных публикаций во многих отраслях науки. Мы не собираемся учить читателя *вычислять* статистические характеристики. Это вполне можно оставить компьютеру.

Читательская аудитория. Эта книга адресована всем гражданам и научным работникам, имеющим желание понять принципы статистики и научиться интерпретировать ее результаты, не вдаваясь в математические детали вычислений. Как ни странно, этой цели можно достичь в очень короткой книжке, содержащей совсем немного уравнений. Мы думаем, что любой человек (а не только студент или ученый), имеющий или не имеющий предварительную подготовку в области статистики, сможет извлечь из этой книги пользу.

Мы свели уровень необходимой математической подготовки к минимуму и всюду, где возможно, обращались к интуиции. Уравнения мы включали только тогда, когда они делали изложение понятнее. Для понимания основных идей достаточно самой элементарной математики и лишь немногих базовых понятий из теории вероятностей. Да и те интуитивно понятны из контекста.

Чего нет в этой книге. Эта книга не курс математической статистики (например, в ней ничего не говорится о борелевской алгебре). Это и не традиционный учебник статистики, в который принято включать многочисленные критерии и методы. Это не руководство по программам статистического анализа типа SPSS или R. Книга не является полным справочником по статистическим критериям. Мы стремились включить достаточно информации для понимания фундаментальных основ статистики, но не больше.

О чем эта книга. В части I мы излагаем философию статистики с минимальным привлечением математики, чтобы были понятны ключевые концепции. Мы познакомимся с самым простым *t*-критерием и покажем, как избежать недоразумений, связанных с вероятностями.

Усвоив материал части I, читатель сможет избежать большинства подводных камней и понять, что на самом деле вычисляют наиболее известные статистические критерии. Мы опишем проверку нулевой гипотезы, не прибегая к сложному математическому аппарату, а применив более простой подход на основе теории обнаружения сигналов (TOC). Материал части II более традиционный, здесь рассматриваются классические критерии: дисперсионный анализ (ANOVA) и корреляция. В частях I и II описаны стандартные статистики – те, что используются наиболее часто. В части III показано, что кризис науки возник из-за неправильного понимания простейших основ статистики, в частности понятия воспроизводимости. Например, читателя, возможно, удивит, что слишком большое число успешных повторений эксперимента может считаться подозрительным явлением, а не свидетельством бесспорности научного факта. Для понимания части III, включающей идеи, которых не найдешь в других вводных учебниках, достаточно лишь базовых понятий и концепций из главы 3 части I. И хотя книга в основном посвящена статистике, мы продемонстрируем тесную связь статистики с планированием эксперимента. Многих статистических проблем можно было бы избежать при выборе правильного – что чаще всего означает «более простого» – плана.

Мы полагаем, что уникальное сочетание базовых концепций статистики (часть I), краткого описания наиболее распространенных статистических критериев (часть II) и нового метастатистического подхода (часть III) позволит не только достичь основательного понимания статистики, но и по-новому – подчас с неожиданной точки зрения – взглянуть на то, что определяет нашу повседневную жизнь.

Материалы. Дополняющие книгу презентации в формате Power Point для преподавателей доступны по запросу на адрес электронной почты michael.herzog@epfl.ch.

Благодарности. Мы благодарны Конраду Нойману и Марку Репнову за корректуру рукописи, а Эдди Кристоферу, Алине Критенуд, Марку, Гертруде и Хайке Херцог, Майе Анне Ястржебовска, Слиму Каммоуну, Иларии Риччи, Эвелине Танелл, Ричарду Уолкеру, Хе Сю и Пьеру Дево за полезные замечания. С грустью сообщаем, что во время подготовки книги к печати от нас ушел Аарон Кларк.

Лозанна, Швейцария — Майкл Х. Херцог
Вест-Лафайетт, Индиана, США — Грегори Фрэнсис
Анкара, Турция — Аарон Кларк

Часть I
Принципы статистики

Глава 1

Основы теории вероятностей

Что вы узнаете из этой главы

Прежде чем ступать на территорию статистики, необходимо вспомнить базовые понятия теории вероятностей. Иначе трудно интерпретировать научные данные, как, впрочем, и информацию в повседневной жизни. В частности, многое из того, что сообщают СМИ, по существу, бесполезно, потому что основано на частичной информации. В этой главе мы объясним, какого рода полная информация нужна для правильных выводов, и познакомимся с теоремой Байеса. Для изложения идей нам понадобятся простые уравнения, а для тех читателей, кто не в ладах с математикой, мы приведем интуитивно понятные соображения на простых примерах и рисунках.

1.1. Путаница вокруг простых понятий теории вероятностей: условные вероятности

1.1.1. Базовый сценарий

Основные понятия теории вероятностей

1. **Вероятность.** Событию A назначается вероятность – число от 0 до 1. Например, при бросании кости вероятность выпадения 4 равна $P(4) = 1/6$.
2. **Распределение вероятностей.** В примере выше имеется 6 возможных исходов, каждому из которых назначена вероятность 1/6. Назначение вероятности каждому возможному исходу дает распределение вероятностей.
3. **Условная вероятность.** Условная вероятность $P(A|B)$ учитывает имеющуюся информацию о событии B. Вертикаль-

ная черта читается «при условии» и означает, что речь идет именно об условной вероятности. Например, мы последовательно вытягиваем две карты из стандартной колоды (52 карты). Вероятность, что первой будет вытянута пиковая масть *P*(первой вытянута пика) = 13/52 = 1/4. Теперь осталась только 51 карта. Вероятность, что во второй раз тоже будет вытянута пика, притом, что она уже была вытянута в первый раз, P(второй вытянута пика|первой вытянута пика) = 12/51. С другой стороны, P(второй вытянута пика|первой вытянута черва) = 13/51. Здесь вероятность вытягивания второй карты зависит от того, какого типа карта была вытянута первой.

4. **Независимые события.** События называются независимыми, если условная вероятность равна безусловной: *P*(A|B) = *P*(A). В этом случае вероятность A не зависит от B. Например, если вытянутая карта возвращается в колоду, то вероятность вытянуть пику во второй раз равна P(второй вытянута пика) = 13/52 вне зависимости от того, какая карта была вытянута первой.

Определения

Рассмотрим ситуацию, когда есть подозрение, что пациент инфицирован, и он сдает соответствующий анализ. Возможны четыре исхода.

1. **Чувствительность:** вероятность положительного анализа при условии, что пациент инфицирован.
2. **Специфичность:** вероятность отрицательного анализа при условии, что пациент не инфицирован.
3. **Частота ложноположительных результатов:** вероятность положительного анализа при условии, что пациент не инфицирован.
4. **Частота ложноотрицательных результатов:** вероятность отрицательного анализа при условии, что пациент инфицирован.

Начнем с примера. В 1980-х годах общество охватила паника: обнаружилась новая болезнь, получившая название «синдром приобретенного иммунодефицита» (СПИД), ее вызывал вирус ВИЧ (*англ.* HIV). Ученые разработали высокочувствительный тест для

определения наличия вируса в крови. Предположим, что и чувствительность, и специфичность теста на ВИЧ равны 0.9999. Это значит, что тест очень хороший, потому что в большинстве случаев он будет давать положительный результат, если пациент инфицирован, и отрицательный, если не инфицирован. Предположим далее, что коэффициент заболеваемости СПИД составляет 0.0001 в нормальной генеральной совокупности, т. е. 1 из 10 000 человек инфицирован вирусом ВИЧ. Анализ, взятый у случайно выбранного человека, оказывается положительным. Допустим, что вы врач. Что вы скажете пациенту о вероятности заболевания? Математически – какова условная вероятность инфицирования ВИЧ при условии, что тест положителен (T^+): $P(HIV|T^+)$?

Поскольку тест очень хорош и почти не дает ошибок, многие полагают, что вероятность $P(HIV|T^+)$ должна быть очень высока, скажем $P(HIV|T^+) = 0.9999$. Однако в действительности $P(HIV|T^+) = 0.5$ – не выше, чем при подбрасывании монеты. Как такое возможно? Мы можем вычислить $P(HIV|T^+)$, воспользовавшись теоремой Байеса, которая ниже сформулирована в общем виде.

Для двух событий A и B

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}.$$

Теперь подставим в эту формулу значения (¬*HIV* означает отсутствие ВИЧ):

$$P(HIV \mid T^+) = \frac{P(T^+ \mid HIV) \times P(HIV)}{P(T^+)} =$$

$$= \frac{P(T^+ \mid HIV) \times P(HIV)}{P(T^+ \mid HIV) \times P(HIV) + P(T^+ \mid \neg HIV) \times P(\neg HIV)} =$$

$$= \frac{0.9999 \times 0.0001}{0.9999 \times 0.0001 + (1 - 0.9999) \times 0.9999} = 0.5.$$

Математика дает ответ, но разобраться в ситуации можно и на интуитивном уровне (рис. 1.1). Предположим, что протестировано 10 000 человек. Поскольку коэффициент заболеваемости равен 0.0001, с высокой вероятностью инфицирован только один из них. Поскольку чувствительность теста очень высока (0.9999), вирус, скорее всего, будет обнаружен. Имеется 9999 неинфицированных людей. Хотя специфичность теста тоже очень высока (0.9999), один ложноположительный результат, вероятно, все же будет – потому

что протестировано много людей. Таким образом, из 10 000 человек всего у двух будут положительные результаты теста (а у 9998 отрицательные). Поскольку из двух человек действительно инфицирован только один, вероятность инфицирования равна ½, т. е. $P(HIV|T^+) = 0.5$.

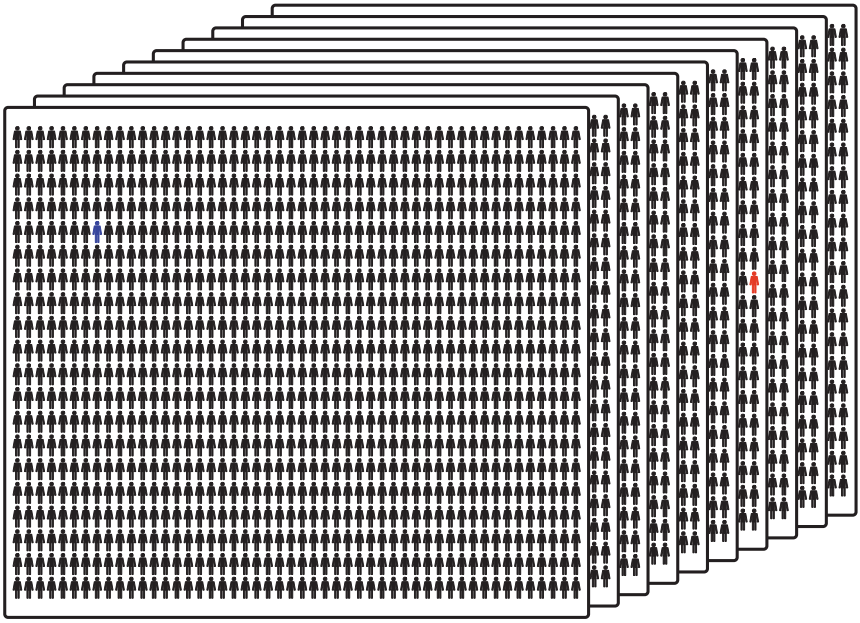


Рис. 1.1. В выборке, включающей 10 000 человек, вероятно, имеется только один инфицированный. Поскольку чувствительность теста высока, результат для этого человека с очень большой вероятностью положителен (выделен красным цветом). Если протестировать одного случайно выбранного неинфицированного человека, то результат с очень большой вероятностью будет отрицательным в силу высокой специфичности теста. Однако всего имеется 9999 неинфицированных людей, и, несмотря на высокую специфичность, тест, вероятно, даст один ложноположительный результат (выделен синим цветом). Итак, мы получили два положительных теста, а поскольку инфицирован только один человек, вероятность инфицирования при положительном тесте составляет 1/2: $P(HIV|T^+) = 0.5$. Очевидно, что игнорировать коэффициент заболеваемость нельзя. Он так же важен, как чувствительность и специфичность

Предположим, что коэффициент заболеваемости еще ниже, например 1/100 000. Допустим, что протестировано 100 000 человек. Поскольку коэффициент заболеваемости равен 1/100 000, вероятно, один из них инфицирован, и тест этого человека, скорее всего,

выявит. Но кроме него на каждые 10 000 человек приходится один ложноотрицательный результат. Следовательно, тест дает 11 положительных результатов[1], и вероятность реального инфицирования при положительном тесте снижается до $P(HIV|T^+) = 1/11 \approx 0.1$. С другой стороны, при коэффициенте заболеваемости 0.5 $P(HIV|T^+) = 0.9999$, т. е. почти 1.0. Следовательно, вероятность $P(HIV|T^+)$ зависит от чувствительности, специфичности *и* коэффициента заболеваемости. Если коэффициент заболеваемости изменяется от 0.0 до 1.0, то $P(HIV|T^+)$ изменяется от 0.0 до 1.0. Для обоснованного заключения необходимо знать все три члена. Если хотя бы один из них неизвестен, все заключения гроша ломаного не стоят. Этот пример иллюстрирует одну из главных тем книги: *помните о частичной информации*!

Замечание 1. Эта демонстрация показывает, насколько важно понимать основы статистических рассуждений. Для пациента умение интерпретировать положительный результат теста имеет огромное значение. Например, в 1987 году 22 реципиента переливания крови получили положительный тест на ВИЧ, и семеро из них покончили жизнь самоубийством [1]. А в одном исследовании отмечается, что в Германии 16 из 20 врачей говорят пациентам, что тест на ВИЧ практически не дает ложноположительных результатов [2].

Замечание 2. Важно, что если вы врач, то ситуация отличается от описанной в примере выше, потому что люди, у которых есть причины подозревать у себя инфекцию, с большей вероятностью сдают тест, чем те, кто более-менее уверен, что не инфицирован. Поэтому коэффициент заболеваемости в больнице, вероятно, выше, чем в примере. Это означает, что $P(HIV|T^+)$ может быть больше 0.5: этот озадачивающий вывод показывает, почему интуитивные представления людей о статистике зачастую не имеют под собой основания.

1.1.2. Второй тест

Правильное понимание природы вероятности также помогает собирать более качественную информацию. Что будет, если мы второй

[1] Мы можем также протестировать 10 000 человек. Поскольку коэффициент заболеваемости равен 1/100 000, с вероятностью 0.1 в этой выборке будет один инфицированный. А при коэффициенте заболеваемости 1/10 000 будет один ложноположительный результат. Поэтому мы должны вычислить частное 0.1/1.1 и получим точно такой же результат – приблизительно 0.1.

раз возьмем тест только у двух человек с положительным первым тестом?[1] Чему теперь будет равна вероятность инфицирования при положительном тесте? Интересующая нас величина равна

$$P(HIV \mid T^{+}) = \frac{0.9999^2 \times 0.0001}{0.9999^2 \times 0.0001 + (1-0.9999)^2 \times 0.9999} = \\ = \frac{0.9999}{0.9999 + 0.0001} = 0.9999.$$

Теперь положительный результат означает, что человек почти наверняка инфицирован.

Это равенство можно объяснить на интуитивном уровне. Первый тест дал два положительных результата. Во второй раз только эти два человека и тестировались. Поскольку тест очень хорош, он почти наверняка определит инфицированного человека и почти наверняка даст отрицательный результат для неинфицированного. Поэтому для человека, сдавшего два положительных теста, вероятность инфицирования близка к 1.0.

Замечание 1. В действительности, назначая повторный тест, врач обнаруживает, что $P(HIV|T^{2+})$ меньше 0.9999. Причина в том, что для некоторых людей тест упорно дает положительный результат, даже если они не инфицированы. Видимо, в их организме имеются молекулы, похожие на антитела, к которым чувствителен тест на ВИЧ.

Замечание 2. Недоразумения вокруг статистики возникают во всех отраслях знания. Корали Колмез и Лейла Шнепс посвятили целую книгу «Math on Trial» анализу судебных дел. В этой книге показано, как непонимание простых положений статистики может привести (и приводило) к неверным выводам. Рассматривается, в частности, дело студентки Аманды Нокс, которая обвинялась в убийстве соседки по квартире. Генетический анализ дал некоторые свидетельства в пользу того, что соседка была убита ножом, на котором присутствовали отпечатки пальцев Аманды. Изучив, какова вероятность точного результата теста, судья решил не проводить повторный тест – несмотря на то, что, как показано выше, повторный анализ мог бы дать совершенно иной результат. Судья был попросту недостаточно знаком с основами статистики [3].

[1] Предполагается, что тесты для данного человека независимы.

1.1.3. Еще пример: синдром Гийена–Барре

Вакцинация (V) от свиного гриппа (SF) может вызвать в качестве побочного эффекта синдром Гийена–Барре (GB) в одном случае на миллион, т. е. $P(GB|V) = 1/1\ 000\ 000$. В тяжелых случаях СГБ напоминает синдром «запертого человека», когда пациент утрачивает способность двигаться и даже разговаривать. Учитывая ужасающие последствия СГБ, стоит ли вообще делать вакцинацию? И на этот раз мы не можем ответить на вопрос, потому что обладаем лишь частичной информацией. Необходимо знать, какова вероятность заболеть синдромом Гийена–Барре без вакцинации ($\neg V$). Предположим, что, помимо вакцинации, СГБ возникает только как осложнение после свиного гриппа (и дополнительно предположим, что вакцина против свиного гриппа стопроцентно эффективна). Вероятность получить СГБ после свиного гриппа довольно высока: 1/3000. Это гораздо больше, чем вероятность 1/1 000 000. Похоже, что вакцинироваться все-таки стоит. Однако следует принять во внимание уровень заражения свиным гриппом, потому что не каждый человек заражается. Этот уровень в каждой новой эпидемии изменяется; предположим, что для случайно выбранного невакцинированного человека вероятность заболеть составляет 1/300. Тогда вероятность получить синдром Гийена–Барре для невакцинированного равна

$$P(GB \mid \neg V) = P(GB \mid SF) \times P(SF \mid \neg V) P(\neg V) = \frac{1}{3000} \times \frac{1}{300} \times 1 = \frac{1}{900\ 000}.$$

Следовательно, в такой ситуации вероятность получить СГБ для невакцинированного лишь ненамного больше, чем для вакцинированного. А вакцина заодно защищает от свиного гриппа.

Важно здесь то, что нельзя принять хорошее решение, основываясь только на одной вероятности (получить синдром Гийена–Барре в результате вакцинации). Нужно также учитывать вероятность дополнительного события (получить синдром Гийена–Барре без вакцинации).

1.2. Недоразумения вокруг вероятностей: отношение шансов

1.2.1. Основные сведения об отношении шансов (ОШ)

Многие курильщики умирают от инфаркта. Бросать курить? Это частичная информация! Встречный вопрос: сколько некурящих уми-

рает от инфаркта? Без этой информации попытка ответить на первый вопрос будет ничуть не лучше утверждения «100 % курильщиков однажды умрут – равно как и 100 % некурящих».

Обобщим эту ситуацию, введя в рассмотрение понятие шанса. Гипотетический пример: из 107 курящих семь перенесли инфаркт, т. е. 100 не переносили (табл. 1.1A). Шансом называется отношение 7/100. Для некурящих инфаркт перенес 1 человек из 100, поэтому шанс равен 1/100. Идея отношения шансов (ОШ) – сравнить две дроби, поделив одну на другую. Это отношение двух отношений говорит нам, в какой степени курящие страдают от инфаркта чаще, чем некурящие: 7/100 / 1/100 = 7. Таким образом, у курящего шанс получить инфаркт в семь раз больше, чем у некурящего, – немало. Для сравнения – если бы никакого эффекта не было, т. е. инфаркт случался бы у курящих и некурящих с одинаковой частотой, то ОШ = 1.0.

Таблица 1.1. Гипотетический пример

A	Курящие	Некурящие		B	Курящие	Некурящие
Инфаркт был	7	1		Инфаркт был	7	1
Инфаркта не было	100	100		Инфаркта не было	10 000	10 000

A) Каковы шансы получить инфаркт, не будучи курильщиком? Предположим, что из 107 курящих семеро перенесли инфаркт, а из 101 некурящего – только один. Насколько шансы курящего получить инфаркт выше, чем у некурящего? Для вычисления отношения шансов мы сначала вычисляем 7/100 и 1/100, а затем делим эти отношения: (7/100) / (1/100) = (7*100) / (1*100) = 7/1= 7. Следовательно, шансы в семь раз выше, что представляется существенным.

B) Теперь предположим, что в группах курящих и некурящих по 10 000 человек, не перенесших инфаркт. Отношение шансов равно (7/10 000)/ (1/10 000) = 7/1 = 7, т. е. не изменилось. Таким образом, отношение шансов не зависит от коэффициента заболеваемости. Однако шанс получить инфаркт уменьшился примерно в 100 раз. Получить ли инфаркт в 7 из 107 случаев или в 7 из 10 007 – «две большие разницы». Отношение шансов дает лишь частичную информацию!

Таблица 1.2. Члены, вносящие вклад в отношение шансов[a]

	С фактором риска	Без фактора риска
Болен	a	b
Не болен	c	d

[a] Небольшое замечание: для вычисления отношения шансов производится деление *a/b* на *c/d*. Можно было бы также воспользоваться пропорциями *a* / (*a* + *b*), *c* / (*c* + *d*) и взять их отношение.

В общем виде (табл. 1.2) отношение шансов $a/c \,/\, b/d = a * d \,/\, b * c$.

Отношение шансов – очень компактный способ сравнения экспериментального и контрольного условия. Оно чаще других показателей используется в медицине и биологии. Например, влияние гена на заболевание обычно выражают в терминах ОШ. Однако решения, принимаемые на базе ОШ, основаны на частичной информации. И вот почему. Увеличим количество людей, перенесших инфаркт, в обеих группах в 100 раз. Отношение шансов при этом не изменится (см. табл. 1.1В).

Очевидно, что отношение шансов не зависит от доли незатронутых людей, пусть даже вероятность получить инфаркт существенно изменилась. Поскольку ОШ не зависит от коэффициента заболеваемости, высокое ОШ почти ничего не говорит, если, к примеру, заболевание редкое.

Как интерпретировать отношения шансов? Во-первых, высокое ОШ – причина для тревоги, только если основной эффект, *a/c*, тоже велик. Например, отношение (число курящих, перенесших инфаркт) / (число курящих, не перенесших инфаркт) = 7/10 000 нельзя считать значительным эффектом, пусть даже ОШ, равное 7, велико. В табл. 1.1В инфаркты просто не очень часты. Только 8 человек из 20 008 перенесли инфаркт. Поэтому крайне маловероятно, что у кого-нибудь вообще был инфаркт, – в отличие от случая в табл. 1.1, где инфаркт перенесли 8 из 208 человек. И в таком случае есть повод для беспокойства. Во-вторых, высокий основной эффект *a/c* – причина беспокоиться, только если ОШ также велико. Вот крайний пример. Если у вас голубые глаза, то вы умрете с очень высокой вероятностью (100 %). Однако для кареглазых вероятность умереть также равна 100 %. Следовательно, ОШ = 1.0, это мало. Можно тревожиться по поводу смерти, но не по поводу цвета глаз.

1.2.2. Частичная информация и мир, полный болезней

Ситуация в целом может быть и еще более запутанной. Мы обсудили влияние одного фактора (курения) на исход (инфаркт). Но курение может также влиять на другие заболевания положительно или отрицательно (даже курение не всегда вредно). Поэтому, чтобы формально ответить на вопрос, следует ли бросать курить, нужно принять во внимание все заболевания, в том числе потенциально неизвестные. Кроме того, нужно учесть затраты на лечение – кариес все же не так серьезен, как инфаркт. Таким образом, нужно вычислить своего рода эффект заболеваемости, который принимает во внимание стоимость различных заболеваний и вероятность их возникновения для данного фактора:

$$\text{Заболеваемость}(\text{Фактор}) = \sum_{S} P(\text{заболевание } S|\text{Фактор}) \times \text{Стоимость}(\text{заболевание } S).$$

Итак, нужно учитывать все болезни, даже еще не открытые. Однако решить, стоит ли бросать курить или менять диету, почти невозможно, если величина эффекта невелика. На практике вся эта информация никогда не бывает доступна. Но это не значит, что статистические рассуждения никогда нельзя использовать для принятия решения. Нужно лишь понимать, что такие решения основаны на неполной информации. И это понимание должно побудить вас собирать как можно больше информации.

Что следует запомнить

1. Не забывайте о частичной информации и старайтесь получить полную информацию, чтобы делать правильные выводы. Например, отношение шансов обычно несет слишком мало информации.
2. Коэффициенты заболеваемости различными болезнями обычно малы, за исключением таких как кариес.

Литература

1. Gigerenzer G, GaissmaierW, Kurz-Milcke E, Schwartz L,Woloshin S. Glaub keiner Statistik, die du nicht verstanden hast. Geist und Gehirn. 2009;10:34–39.
2. Gigerenzer G, Hoffrage U, Ebert A. AIDS counselling for low-risk clients. AIDS Care. 1998;10:197–211.
3. Colmez C, Schneps L. Math on trial: how numbers get used and abused in the courtroom. Basic Books: New York; 2013.

Глава 2

Планирование эксперимента и основы статистики: теория обнаружения сигналов (ТОС)

Что вы узнаете из этой главы

Что считать хорошей мерой качества? Чаще всего используется процентная доля правильных ответов. Ниже мы увидим, что в этом показателе смешиваются две переменные, – чувствительность и порог, – поэтому относиться к ней следует с осторожностью. Мы введем меру чувствительности d', которая оказывается исключительно важной во многих областях статистики.

2.1. Классический сценарий ТОС

Представим, что мы находимся в желтой подводной лодке, пересекающей океан. Было бы очень опасно столкнуться со скалой, поэтому подлодка оборудована гидролокатором. Гидролокатор излучает волны, а приемник получает их отражения. Эти отраженные волны объединяются в «гидроакустическую характеристику». Если имеется скала, то гидроакустическая характеристика будет больше, чем в случае отсутствия скалы. Однако картину искажают шумы, поэтому даже при одинаковых объективных условиях – скала есть или скалы нет – регистрируемая гидроакустическая характеристика заметно отличается (рис. 2.1).

Каждому из двух возможных условий соответствует распределение вероятностей, показывающее, насколько вероятно некоторое значение гидроакустической характеристики на оси x.[1] Как часто бывает

[1] Строго говоря, вероятность, что гидроакустическая характеристика *в точности* равна 80, нулевая. Функция вероятности показывает, что наблюдается значение, близкое к 80 (находящееся в очень узком интервале, окружающем 80). Для математической статистики такие детали крайне важны, но для понимания основ статистики роли не играют.

в статистике, мы предполагаем, что нормальное (гауссово) распределение адекватно описывает ситуацию. Нормальное распределение полностью определено своим средним значением μ и стандартным отклонением σ, определяющим ширину гауссианы. Большое значение σ означает, что для одного и того же условия присутствия скалы отраженные сигналы с высокой вероятностью будут различаться. Наоборот, если $\sigma = 0$, то никакой изменчивости нет вообще, т. е. в качестве гидроакустической характеристики мы всегда получаем одно и то же значение. Поэтому σ отражает зашумленность и так и называется – шум. Отсутствие скалы мы называем *чистым шумом*, поскольку сигнал не отражается от скалы, и присутствие скалы – *сигнал плюс шум*.

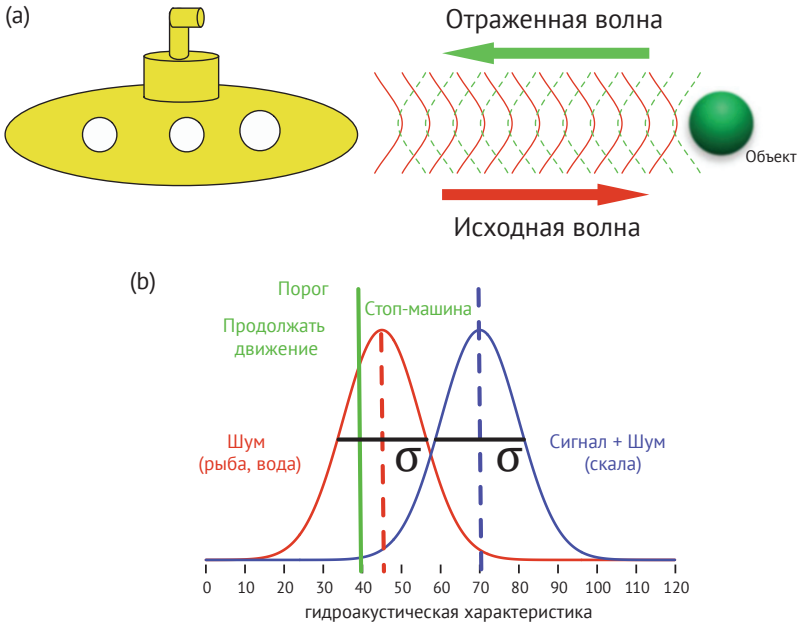


Рис. 2.1 (**a**) Подлодка посылает гидроакустические сигналы и измеряет отраженные сигналы. Гидроакустическая характеристика зашумлена, т. е. для одной и той же скалы отраженные сигналы в разных измерениях могут отличаться. Например, на отражающую способность может повлиять проплывающая рыба. То же самое справедливо, когда никакой скалы нет. Как решить, есть скала или нет? (**b**) Классический сценарий ТОС. Оба условия – присутствие и отсутствие скалы – принадлежат какому-то распределению вероятностей, которое показывает, насколько вероятно получение различных значений гидроакустической характеристики. ТОС предполагает, что эти распределения вероятностей нормальные

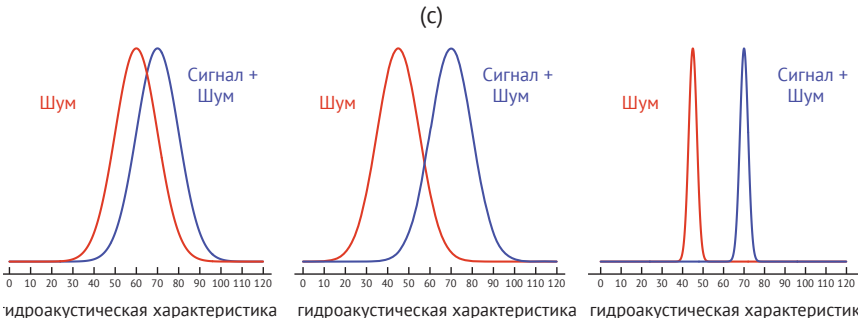


Рис. 2.1 (*Окончание*). В этом примере среднее значение гидроакустической характеристики равно 70, когда скала присутствует, и 45, когда ее нет. Размах распределения описывается стандартным отклонением σ. Здесь σ = 10. Если регистрируется гидроакустическая характеристика 80, то присутствие скалы гораздо вероятнее отсутствия. Характеристика 57.5 с одинаковой вероятностью может описывать как присутствие, так и отсутствие скалы. Чтобы принять решение, необходим порог. Если гидроакустическая характеристика больше порога, мы принимаем решение остановить двигатель (потому что впереди скала), в противном случае продолжаем движение (скалы нет). Как установить порог, решаем мы сами. Хотим осторожничать – выберем порог поменьше, хотим рисковать – побольше. (**c**) Насколько точно мы можем отличить присутствие скалы от отсутствия, зависит от перекрытия двух распределений вероятностей. Перекрытие, обозначаемое символом d', – это разность между средними значениями, поделенная на стандартное отклонение. Большое перекрытие означает, что способность различать условия низкая, а малое – что высокая. При фиксированном стандартном отклонении d' растет по мере роста разности средних (ср. левый и средний рисунок). При фиксированной разности средних d' растет по мере убывания стандартного отклонения (ср. средний и правый рисунок)

Насколько хорошо мы можем отличить присутствие скалы от отсутствия? Зависит от перекрытия гауссиан. Если гауссианы перекрываются на 100 %, то различить эти два случая невозможно, и гидролокатор бесполезен. Если перекрытия почти нет, то различить ситуации легко, потому что данное значение гидроакустической характеристики с большой вероятностью принадлежит только одной гауссиане. Например, на рис. 2.1b значение гидроакустической характеристики 80 соответствует присутствию скалы с гораздо большей вероятностью, чем ее отсутствию. Перекрытие можно оценить разностью между средними μ_1 и μ_2 обеих гауссиан,

поделенной на стандартное отклонение σ, в предположении, что σ для обоих распределений одинаково[1]:

$$d' = \frac{\mu_1 - \mu_2}{\sigma}. \tag{2.2}$$

d' называется чувствительностью или различимостью и измеряет, насколько хорошо можно различить две альтернативы, т. е. d' – мера отношения сигнала ($\mu_1 - \mu_2$) к шуму (σ). Перекрытие зависит как от разности средних, так и от стандартного отклонения. Следовательно, увеличить d' можно двумя способами: увеличив разность средних или уменьшив стандартное отклонение (рис. 2.1c). Важно, что понятие чувствительности здесь не то же самое, что одноименное понятие, введенное в главе 1, которое можно иначе назвать частотой истинно положительных результатов. Далее мы будем употреблять термин «чувствительность» только в этом последнем смысле, но не для обозначения d'.

Когда следует останавливать двигатель? Необходим порог принятия решения *c*. На рис. 2.1b выбран порог 40: если гидроакустическая характеристика больше 40, мы останавливаем двигатель, иначе продолжаем движение. Какое значение порога выбрать, зависит от нас. Если присутствие скалы так же вероятно, как ее отсутствие, то оптимальным порогом будет точка пересечения двух гауссиан, поскольку при этом достигается максимум количества правильных решений.

2.2. TOC и доля правильных ответов

Применим TOC к типичному поведенческому эксперименту, связанному с обнаружением сигнала. Вы смотрите на экран компьютера и видите либо тусклое световое пятнышко (стимул присутствует), либо пустой экран (стимул отсутствует). Как и в примерах с желтой подводной лодкой и тестом на ВИЧ, имеется четыре возможных исхода (табл. 2.1).

---

[1] В теории обнаружения сигналов (TOC) d' обычно определяется с использованием абсолютной величины разности средних:

$$d' = \left| \frac{\mu_1 - \mu_2}{\sigma} \right|. \tag{2.1}$$

Мы не стали использовать абсолютную величину, потому что такое определение удобнее для применения d' в статистике в главе 3.

Если стимул присутствует в половине испытаний, то процентная доля правильных ответов вычисляется как среднее между частотой правильных подтверждений (Hit) и частотой правильных пропусков (Correct Rejection – CR): (*Hit* + *CR*) / 2. Рассмотрим «процент правильных» в терминах TOC. Как и в примере с подлодкой, предполагается зашумленность. Но в отличие от этого примера мы не знаем, как процессы восприятия кодируются в мозге человека, т. е. не имеем явных распределений вероятностей. Зато у TOC есть важное достоинство – большая гибкость. Например, предположим, что можно сосредоточить все внимание на одном нейроне, который кодирует яркость стимула. Значение 0.0 соответствует пустому экрану, а положительное значение – световому пятнышку определенной яркости. Будем использовать порог принятия решения, который определяет, что распознать: световое пятно или пустой экран. При консервативном поведении порог задается более высоким, т. е. мы отвечаем «световое пятно есть», только когда уверены в этом. Если мы готовы рисковать, то значение порога задается меньшим, т. е. мы подтверждаем наличие светового пятна, когда есть малейшие признаки света. Мы можем задать любое значение порога. Например, чтобы оптимизировать процент правильных ответов, можно выбрать в качестве порога пересечение гауссиан. В этом случае мы даем равное число ответов ««световое пятно есть» и «светового пятна нет», если оба варианта стимула предъявляются с одинаковой частотой. Если пятно чаще отсутствует, чем присутствует, то имеет смысл сдвинуть порог в сторону ответов «светового пятна нет» – в данном случае вправо. Кроме того, можно принять во внимание цену ответа. Если вознаграждение за ответ «световое пятно есть» меньше, чем за ответ «светового пятно нет», то лучше сдвинуть порог так, чтобы ответ «световое пятно есть» был более частым.

Таблица 2.1. Четыре исхода в классическом эксперименте TOC

Ответ	Стимул присутствует	Стимул отсутствует
Присутствует	Правильное подтверждение (Hit)	Ложная тревога (False Alarm – FA)
Отсутствует	Ошибочный пропуск (Miss)	Правильный пропуск (Correct Rejection – CR)

Скала (световое пятно) присутствует, и мы ответили, что присутствует (правильное подтверждение). Скала (световое пятно) отсутствует, а мы ответили, что присутствует (ошибочный пропуск). Скала (световое пятно) отсутствует, а мы ответили, что присутствует (ложная тревога). Скала (световое пятно) отсутствует, и мы ответили, что отсутствует (правильный пропуск).

Давайте будем гладко изменять порог и посмотрим, как меняется процент правильных ответов в зависимости от d' (рис. 2.2). Начнем с порога 2.0, т. е. точки пересечения гауссиан. Если немного сдвинуть порог в сторону от оптимума, например вправо, то процент правильных ответов уменьшается, и тем больше, чем дальше мы отодвигаем порог. Если сдвинуть порог слишком далеко вправо, то качество стремится к 50 %, т. е. к уровню случайного угадывания. В этом случае мы всегда даем один и тот же ответ, например «световое пятно есть», т. е. велико *смещение ответа*. Важно, что различимость d' при этом не изменилась, т. е. наша способность к зрительному распознаванию осталась постоянной. Изменилось только наше поведение в части принятия решений.

Таким образом, процент верных ответов смешивает в одно порог принятия решения и различимость. Для определенного уровня качества, скажем 75 %, мы не можем сказать, что имеет место: высокая различимость и неоптимальный порог или оптимальный порог и низкая различимость. Поэтому процент верных ответов может оказаться опасным показателем, который легко способен затушевать истинные эффекты или давать ложноположительные результаты. Выводы, базирующиеся на проценте верных ответов, основаны на частичной информации!

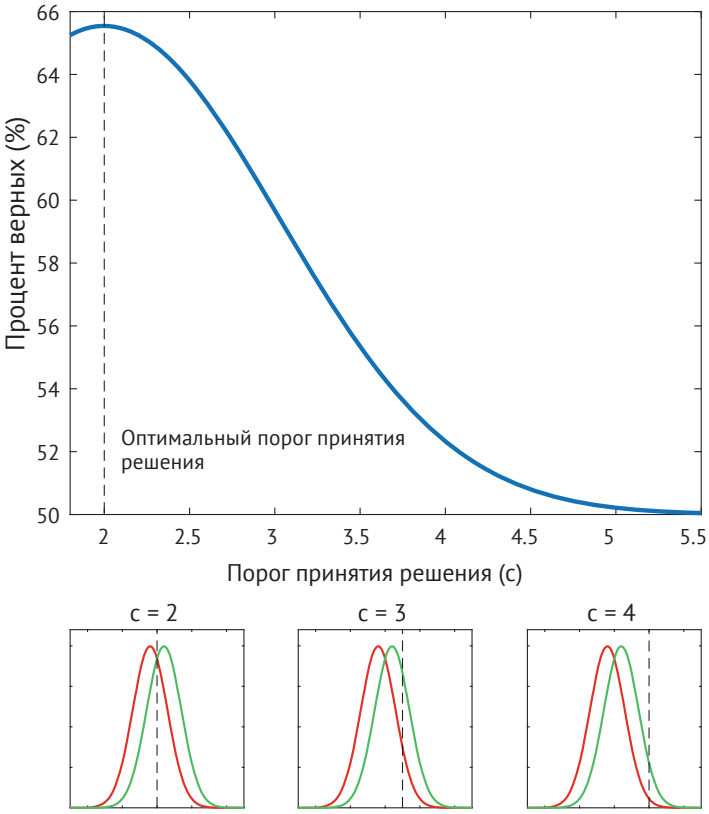


Рис. 2.2. Процент верных ответов зависит от порога. Сначала поместим порог в точку пересечения двух гауссиан, т. е. положим равным 2 (левый нижний рисунок). Затем будем сдвигать порог вправо (средний и правый рисунок). На верхнем рисунке показано, что процент правильных ответов уменьшается по мере отодвигания порога от оптимума. Если порог сдвинут очень далеко вправо, то мы почти всегда даем один и тот же ответ, например «светового пятна нет». В этом случае частота правильных подтверждений равна 0, а частота ошибочных пропусков равна 1. Следовательно, процент верных ответов равен (0 + 1) / 2 = 0.5, что совпадает со случайным угадыванием, хотя, в принципе, мы способны различить обе ситуации

2.3. Эмпирическая d'

Зная только процент верных ответов, невозможно вывести различимость из эксперимента. Значит ли это, что все эксперименты безнадежны? Как ни удивительно, мы можем разделить различимость и порог, если будем по отдельности оценивать *d'* и *b*, *смещение* порога:

$$d'_{emp} = z(Hit) - z(FA), \tag{2.3}$$

$$b_{emp} = \frac{z(Hit) - z(FA)}{2}. \tag{2.4}$$

Для вычисления d' нам просто нужно применить z-преобразование к частотам правильных подтверждений и ложных тревог. z-преобразованием в данном случае является обратная гауссова функция распределения. Если вы незнакомы с z-преобразованием, просто рассматривайте его как функцию, которую можно найти в библиотеках на вашем компьютере. b_{emp} говорит, насколько текущий порог отличается от оптимального, т. е. как далеко он отстоит от точки пересечения гауссиан. Следовательно, b_{emp} измеряет смещение порога.

Важно, что d'_{emp} не изменяется при изменении порога (и, стало быть, смещении ответа). Однако, не завися от порога, d' все же зависит от модели. Предполагается, что выполнено три условия:

1) истинные распределения вероятностей нормальные,
2) эти нормальные распределения имеют одинаковую дисперсию,
3) в процессе измерений порог не изменяется.

Условие 1 принципиально важно, потому что мы вычисляем z-преобразование, т. е. обратное нормальное распределение, а это имеет смысл, только когда распределение данных нормальное. Условие 1 часто выполняется. Условие 2 обычно выполняется, потому что альтернативные стимулы похожи. Условие 3 очень важно, но проверить его нелегко.

Внимание

Термин «чувствительность» употребляется в двух разных смыслах.

1. Преимущественно в медицинской литературе чувствительность – то же самое, что частота правильных подтверждений (Hit Rate).
2. В TOC чувствительность соответствует различимости, т. е. d' $z(Hit) - z(FA)$. В контексте TOC мы будем употреблять термин «различимость», а не «чувствительность».

Пример 1 (автоматизированная система). На рис. 2.3 показано качество (частоты правильных подтверждений, ложных тревог, ошибочных пропусков и правильных пропусков) работы врача и системы на основе искусственного интеллекта (ИИ)

при диагностике заболевания. Общий процент правильных ответов в обоих случаях равен 80 %. Вычислить d' почти так же просто, как процент верных ответов: нужно только применить z-преобразование к частотам правильных подтверждений и ложных тревог. Как выясняется, в терминах d', в отличие от процента верных, качество сильно различается. Кто лучше: врач или система ИИ? Обычно различимость – встроенная характеристика системы, изменить которую трудно. Ваши глаза такие, как есть, – не лучше и не хуже. С другой стороны, изменить порог принятия решения просто, нужно лишь чаще выбирать один ответ в ущерб другому. Очевидно, что система ИИ сильно смещена в сторону ответов «да», что помогает избежать ошибочных пропусков (ложноотрицательных результатов), но увеличивает частоту ложных тревог (ложноположительных результатов). Поэтому порог системы ИИ далек от оптимального. Задание оптимального порога существенно повышает качество в терминах процента верных ответов.

Качество работы врача			**Автоматическое распознавание**		
Сигнал	**Есть**	**Нет**	**Сигнал**	**Есть**	**Нет**
Да	80	20	Да	98	38
Нет	20	20	Нет	2	62
	P	*z*		**P**	*z*
Hit	0.8	0.842	Hit	0.98	2.054
FA	0.2	−0.842	FA	0.38	−0.305
Чувствительность d'	1.683		Чувствительность d'	2.359	
Смещение b	0.000		Смещение b	−0.874	
P(правильно)	0.800		P(правильно)	0.800	

Рис. 2.3. Сравнение врача с машиной. Процент верных ответов в обоих случаях одинаков. Вычислить d' почти так же просто, как процент верных ответов: нужно только применить z-преобразование к частотам правильных подтверждений и ложных тревог и найти разность. В терминах d' качество, т. е. различимость, сильно отличается. Кто лучше: врач или система ИИ? Очевидно, что система ИИ сильно смещена в сторону ответов «да», что помогает избежать ошибочных пропусков, но увеличивает частоту ложных тревог. Порог системы ИИ далек от оптимального. Печатается с разрешения Марка Джорджсона

Пример 2 *(обучение)*. В примере обучения наблюдателям предъявляется отрезок прямой, немного наклоненный вправо или влево. Поскольку различие мало, наблюдатели часто ошибаются. Для повышения качества наблюдатели обучаются на задаче с 10 блоками, каждый из которых содержит 80 испытаний. Количество верных ответов в каждом из 10 блоков усредняется по всем 80 испытаниям. Качество резко возрастает. Означает ли это, что улучшилось восприятие? Улучшение может быть вызвано изменением различимости или порога. Например, обучение со стимулами может приводить к уменьшению дисперсии σ гауссиан, т. е. люди начинают более точно различать наклон линий. Уменьшение дисперсии ведет к увеличению d', т. е. к увеличению различимости (рис. 2.2). Увеличение различимости может также иметь место, если раздвинуть средние гауссиан. Качество может также улучшиться, когда в блоке 1 порог принятия решений участниками не оптимален, а в процессе обучения участники, возможно, подправляют порог. Изменение восприятия обычно считают связанным с изменением различимости. Применяя для анализа данных процент верных ответов, мы не можем прийти к правильным выводам, потому что не в состоянии отделить изменения различимости от изменений порога. Поэтому во всех экспериментах по обучению важно строить графики зависимости результатов от d' и смещения.

Пример 3 *(чувствительность и специфичность)*. Тест на ВИЧ, рассматриваемый в главе 1, имел очень высокую чувствительность и специфичность. Определение чувствительности и специфичности зависит также от порога. На самом деле оно ничем не отличается от определения процента верных. Напомним, что чувствительность – это частота правильных подтверждений (истинно положительных результатов), а специфичность – частота правильных пропусков (истинно отрицательных результатов). Поэтому ситуация в точности такая же, как в приведенном выше примере с подводной лодкой, только теперь по оси x откладывается концентрация антител к ВИЧ (измеряемая тестом). Нам необходим порог, который определяет, является ли тест положительным для некоторой концентрации антител. Следовательно, мы можем увеличивать чувствительность за счет специфичности и наоборот. Просто для справки отметим, что (чувствительность + специфичность) / 2 – это процент верных ответов.

Пример 4 *(компромисс между скоростью и точностью).* Во многих экспериментах уменьшают время, отведенное на ответ, т. е. заставляют наблюдателей отвечать как можно быстрее. Зачастую медленные наблюдатели, например пожилые люди, демонстрируют более высокую d', чем быстрые (более молодые) наблюдатели; это так называемый компромисс между скоростью и точностью. Таким образом, ситуация еще усложняется, поскольку необходимо сопоставлять время реакции с d' и смещением, чтобы прийти к правильным выводам. Эксперименты с четко выраженным компромиссом между скоростью и точностью часто с трудом поддаются интерпретации.

Пример 5 *(эффекты пола и потолка).* Еще одна проблема связана с так называемыми эффектами пола и потолка. В эксперименте (бессмысленном) экспериментатор поднимает руку, растопырив пальцы. Все наблюдатели правильно идентифицируют все пять пальцев, т. е. число верных ответов 100 %. Можно ли отсюда сделать вывод, что у всех наблюдателей одинаково хорошее зрение, т. е. способность к различению? Конечно, нет; задача была слишком простой, поэтому наблюдатели находились в режиме потолка, когда качество близко к 100 %. Делать какие-либо выводы бесполезно. Вычисление d' в этой ситуации не поможет, потому что частота ложных тревог равна 0.0 и d' равно бесконечности.

То же самое справедливо для эффекта пола, когда качество близко к уровню случайного гадания (50 %). Поэтому важно следить за тем, чтобы альтернативные стимулы находились в диапазоне, позволяющем обнаружить различия между участниками.

Пример 6 *(стандартизованные эффекты).* d' часто называют стандартизованным эффектом, потому что деление на σ приводит измерения к единицам стандартного отклонения. В результате d' становится нечувствительно к оригинальным единицам измерения (т. е. неважно, производились ли оригинальные измерения в метрах или в дюймах). Кроме того, размер стандартизованного эффекта часто может быть нечувствителен к некоторым вариациями эксперимента. Например, если экспериментом на время реакции можно манипулировать, так чтобы замедлить все действия в одно и то же число раз, то разность средних (сигнал) увеличится, а стандартное отклонение (шум) уменьшится в одно и то же число раз. Отношение же d' останется неизменным.

Что следует запомнить

1. Не забывайте о частичной информации. Процент верных ответов смешивает в одно различимость d' и порог принятий решений c.
2. Не забывайте о частичной информации. Это относится и ко многим другим показателям, например чувствительности и специфичности в медицинских тестах.
3. Разделить различимость и порог поможет d'_{emp}.
4. d'_{emp} не зависит от порога, но зависит от модели. Бесплатных завтраков не бывает.

Глава 3

Главная концепция статистики

Что вы узнаете из этой главы

В главах 1 и 2 мы показали, что для обоснованных выводов необходима полная информация и что многие популярные показатели, например отношение шансов или процент верных ответов, дают лишь частичную информацию. В этой главе мы воспользуемся идеями ТОС, чтобы разобраться в статистическом выводе, включая роль *p*-значений, встречающихся в статистике сплошь и рядом. Мы покажем, что *p*-значение смешивает в одно размер эффекта и размер выборки, а следовательно, также дает лишь частичную информацию.

Эта глава посвящена сущностным основам статистики. Мы объясним их на примере *t*-критерия, самого простого и популярного статистического критерия. Это единственная глава книги, в которой мы опускаемся на уровень деталей, потому что, на наш взгляд, эти детали очень способствуют пониманию фундаментальных аспектов статистики. Но все равно понадобится лишь элементарная математика. Читатель, который торопится или испытывает непреодолимое отвращение к математике, может перейти сразу к разделу 3.3 «Резюме», где мы перечислим главные факты и основные шаги. Для дальнейшего чтения необходимо понимать хотя бы это резюме.

3.1. Еще один способ оценки отношения сигнал–шум

В главе 2 мы определили d' как расстояние между средними распределений двух генеральных совокупностей, поделенное на стандартное отклонение. Обычно одну из генеральных совокупностей мы называем чистым шумом со средним μ_N, а другую – зашумленным сигналом со средним μ_{SN}. В статистике d' генеральных совокупностей часто называют дельтой (δ) Коэна или размером эффекта. Вычисляется она так же, как в главе 2:

$$\delta = d' = \frac{\mu_{SN} - \mu_N}{\sigma}. \tag{3.1}$$

Часто у нас нет информации о генеральной совокупности, и мы хотим оценить δ (т. е. d') по эмпирическим данным. Например, в главе 2 мы могли бы оценить d' в эксперименте с присутствием или отсутствием светового пятна, вычислив z(Hit) – z(FA) по одним лишь данным о поведении. Мы ничего не знали о средних и дисперсии нормальных распределений. Такой подход на основе оценивания полезен, когда мы не можем непосредственно измерить величины, определяющие качество системы, но можем измерить последствия принятых решений. Бывает и так, что измерить последствия решений трудно, но можно непосредственно оценить средние и дисперсию нормальных распределений. Например, мы можем использовать гидролокатор и регистрировать гидроакустическую характеристику во многих испытаниях, когда скала присутствует. Затем наносим результаты на график, по которому можно оценить среднее и дисперсию. Можно таким же образом получить среднее и дисперсию для случая, когда скала отсутствует. Затем выборочные средние $\bar{x}_{SN}$ и $\bar{x}_N$ и стандартное отклонение s можно использовать для вычисления оценки размера эффекта, называемой d Коэна:

$$d = \frac{\bar{x}_{SN} - \bar{x}_N}{s}. \tag{3.2}$$

И снова этот стандартизованный эффект d является просто оценкой d' распределений генеральной совокупности. Если d велико, то будет достаточно просто отнести одиночное измерение к распределению зашумленного сигнала или к распределению чистого шума. К сожалению, во многих ситуациях, представляющих интерес для ученых, величина d очень мала. Например, в психологии значение в районе $d = 0.8$ считается «большим», в районе $d = 0.5$ – «средним», а в районе $d = 0.2$ – «малым». Как показано на рис. 3.1, даже «большие» значения d соответствуют значительному перекрытию распределений. В таких случаях мы никогда не добьемся приемлемой способности правильно различать *одиночные* измерения. Но не все потеряно, коль скоро мы готовы удовлетвориться различением *средних* измерений. Как мы увидим, средства ТОС применимы для различения средних подобно тому, как различаются одиночные измерения.

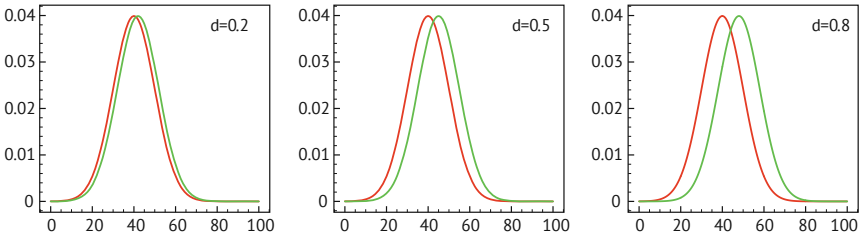


Рис. 3.1. Распределения генеральной совокупности с малым (d = 0.2), средним (d = 0.5) и большим (d = 0.8) размером эффекта

Терминология

Поскольку многие похожие понятия возникали в разных областях знания, для них употребляются разные термины. Некоторые из них перечислены ниже.

- Частота правильных подтверждений = Мощность.
- Частота ложноположительных результатов = Ложная тревога = Ошибка типа I.
- Частота ошибочных пропусков = Ошибка типа II.
- d' = δ Коэна = Размер эффекта = Стандартизованный размер эффекта.
- Гауссово распределение = Нормальное распределение = Колоколообразная кривая.
- Выборочные значения, такие как высота дерева, называются также отметками (score).

Некоторые определения

Допустим, имеется выборка n оценок x_i. Тогда определены следующие величины.

- Выборочное среднее:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

где символ $\sum$ означает «сумма всех последующих членов».

- Выборочная дисперсия:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

- Выборочное стандартное отклонение:

$$s = \sqrt{s^2}.$$

- Стандартная ошибка:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}.$$

Свойства выборочных средних

При $n \to \infty$:

1) распределение выборочных средних $\bar{x}$ является нормальным (центральная предельная теорема – ЦПТ),

2) $\bar{x} \to \mu$,

3) $s_{\bar{x}} \to 0$.

3.2. Недостаточная выборка

Начнем с примера. Нас интересует следующая гипотеза: средняя высота дубов на северном склоне Альп отличается от средней высоты дубов на южном склоне. Прямолинейный способ проверить эту гипотезу – измерить высоты *всех* деревьев на северном и южном склоне, вычислить средние и сравнить их. Если средние различны, значит, различны. Если одинаковы, значит, одинаковы. Все просто.

К сожалению, дубов очень много (рис. 3.2), а наши возможности ограничены, поэтому мы можем обмерить только определенное количество деревьев, скажем n, на северном и южном склоне. Выбранные из генеральной совокупности деревья называются *выборкой*, т. е. в данном случае мы набрали две выборки: одну с южного склона, другую с северного. Измеренные нами высоты деревьев называются *примерами*, или *образцами*, а средняя высота в выборке – *выборочным средним*. Для каждой выборки *средняя* высота *выборочных* деревьев, скорее всего, отличается от *истинной* высоты *всех* деревьев в генеральной совокупности. Например, по чистой случайности вы-

соких деревьев в выборке могло бы оказаться больше, чем низких. Разность между тем и другим называется *выборочной ошибкой*. Таким образом, из-за *недостаточности выборки* (обмерены не все деревья) мы, скорее всего, не получим точные оценки обоих средних. Важно, что деревья выбираются случайно, а процедура называется *случайным отбором*.

Рис. 3.2. Небольшой участок леса в Швейцарских Альпах

Итак, мы сформировали выборки деревьев с северного и с южного склона. Если обнаружится различие между выборочными средними, то мы не сможем сказать, является ли его причиной действительное различие в генеральных совокупностях деревьев, или истинные средние одинаковы, а все дело в недостаточности выборки. Следовательно, недостаточность выборки может привести к неверным выводам. Например, несмотря на то что средние генеральных совокупностей северного и южного склонов одинаковы, мы можем решить, что разница есть, потому что различаются выборочные средние. В таком случае имеет место ложная тревога, или ошибка типа I.

Чтобы понять, как недостаточность выборки влияет на решения, изучим, во-первых, насколько вероятно, что выборочное среднее

отклоняется от истинного на определенную величину. Как мы увидим, выборочная ошибка определяется стандартным отклонением генеральной совокупности, σ, и размером выборки, n. Во-вторых, мы изучим, как недостаточность выборки влияет на нашу способность определять, существует или не существует различие в средней высоте *двух* генеральных совокупностей деревьев. Ответ выражается простой формулой, которая есть не что иное, как d для средних значений. Таким образом, мы оказываемся в ситуации ТОС. В-третьих, мы хотим контролировать частоту ошибок типа I. Мы увидим, что знаменитое p-значение просто задает порог для ошибок типа I.

3.2.1. Выборочное распределение среднего

Для начала сосредоточимся на генеральной совокупности деревьев северного склона. Чтобы получить выборочное среднее, сформируем выборку деревьев, измерим высоту x_i каждого дерева, сложим все высоты и разделим сумму на размер выборки n. Выборочное среднее – это оценка истинного среднего μ_{North}:

$$\bar{x}_{North} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3.3}$$

где x_i – высота i-го дерева в выборке. Аналогично можно оценить дисперсию высот деревьев, s_2 (см. врезку). Разность между выборочным и истинным средним равна:

$$\bar{x}_{North} - \mu_{North}. \tag{3.4}$$

Насколько велика эта разность *в среднем*? Чтобы ответить на этот вопрос, предположим – для определенности, – что мы много раз ходили в лес и случайным образом набирали выборку фиксированного размера n. В разных выборках, скорее всего, окажутся разные деревья, потому что мы действовали случайным образом. Сколько выборочных средних близки к истинному, а сколько далеки от него? Предположим, что нам известно истинное среднее и стандартное отклонение генеральной совокупности. Сначала включим в выборку только два дерева и вычислим среднее. В примере на рис. 3.3 истинное среднее равно 20 м, а среднее по выборке – 19 м. Таким образом, разность составляет 1 м. Сформируем еще одну выборку из двух деревьев. Ошибка, скорее всего, будет другой, потому что высоты деревьев, надо полагать, отличаются от предыдущих. Продолжим измерять пары деревьев и посмотрим,

как распределены выборочные средние. Согласно центральной предельной теореме распределение выборочных средних похоже на нормальное. Нормальное распределение центрировано вокруг истинного среднего, поэтому при наличии большого числа выборочных средних мы получим хорошее представление об истинном. Однако гауссиана получается довольно широкой, т. е. стандартное отклонение велико, а потому некоторые выборочные средние значительно отклоняются от истинного.

Теперь будем формировать выборки размера 9, а не 2 и повторим всю процедуру. Распределение снова будет нормальным, но более узким, чем в предыдущем случае, т. е. стандартное отклонение стало меньше, а значит, стало гораздо менее вероятно, что среднее случайной выборки сильно отклонится от истинного среднего. Вообще, для каждого размера выборки n существует выборочное распределение. Чем больше n, тем меньше стандартное отклонение выборочного распределения. Это неудивительно, потому что ошибка будет равна нулю, если обмерить все деревья, и будет мала, если количество необмеренных деревьев невелико. Поэтому стандартное отклонение $\sigma_{\bar{x}}$ выборочных распределений отражает наши ожидания относительно качества оценки среднего. Величина $\sigma_{\bar{x}}$ называется *стандартной ошибкой среднего*, и можно показать, что $\sigma_{\bar{x}}$ равна стандартному отклонению истинного распределения генеральной совокупности σ, поделенному на квадратный корень из размера выборки n:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}. \tag{3.5}$$

Если мы не знаем σ, то можем оценить стандартную ошибку, воспользовавшись выборочной оценкой стандартного отклонения:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}. \tag{3.6}$$

Эта формула еще раз показывает, почему при увеличении размера выборки выборочная ошибка уменьшается: когда $\sqrt{n}$ возрастает, $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ стремится к нулю. $\sigma_{\bar{x}}$ зависит от n и σ. Предположим, что σ равно нулю, тогда все деревья в выборке имеют одинаковую высоту, равную средней высоте μ_{North}, и, следовательно, нужно измерить высоту только одного дерева. С другой стороны, если σ велико, то нужно выбрать много деревьев, чтобы получить хорошую оценку среднего генеральной совокупности.

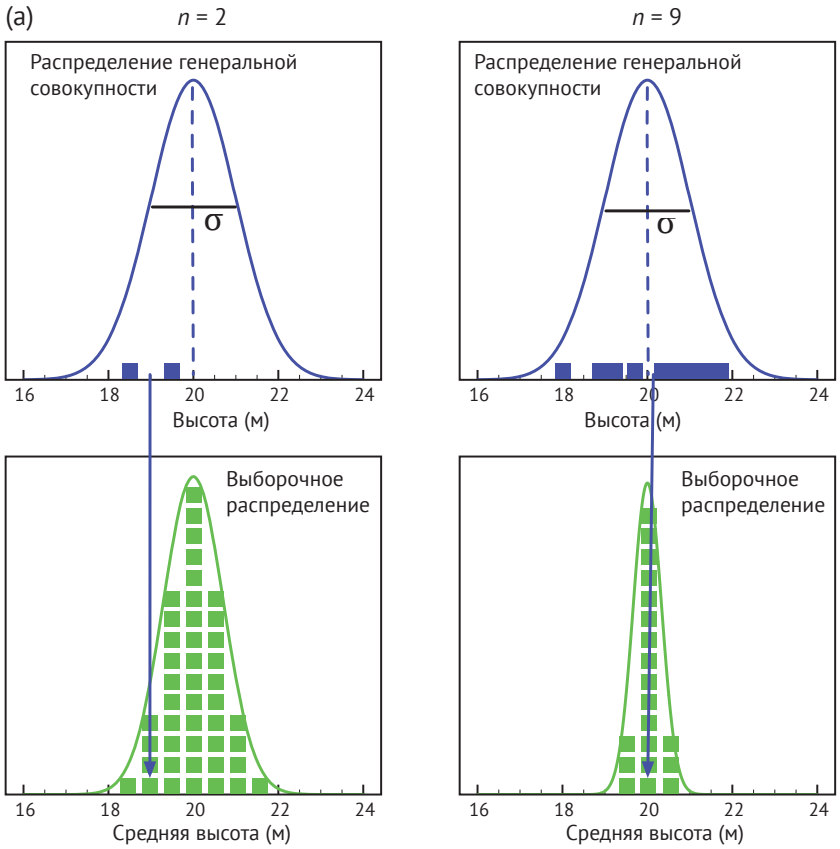


Рис. 3.3. Сосредоточимся только на деревьях северного склона. Мы не можем измерить высоты всех деревьев, поэтому набираем выборки, средние по которым, в силу недостаточности выборки, скорее всего, будут отличаться от истинного среднего. (а, слева) Гауссиана в верхнем ряду показывает истинное распределение генеральной совокупности. Сначала измеряем высоты только двух случайно выбранных деревьев (синие квадратики). Они равны 18.5 и 19.5 м. Следовательно, среднее по выборке размера два равно 19 м, оно показано зеленым квадратиком на графике снизу, на который указывает стрелка. Стало быть, выборочная ошибка среднего равна 1 м, потому что истинное среднее равно 20 м. Сходим в лес и обмерим еще два дерева. Каждый зеленый квадратик соответствует одному из выборочных средних. Набрав много таких выборок, мы получим гауссову функцию. По оси *y* отложена вероятность появления некоторого среднего значения. (а, справа) Пусть теперь выборка состоит из 9 деревьев. Истинное распределение генеральной совокупности такое же, как на левом рисунке, оно показано в верхнем ряду на примере одной выборки размера 9. Соответствующее выборочное среднее обозначено стрелкой. И снова дополнительные выборки дают нам новые средние значения (зеленые квадратики). Гауссиана стала уже, т. е. стандартное отклонение выборочных средних меньше. Таким образом, для выборки из двух деревьев ошибка 2 м вполне вероятна, но для выборки из девяти деревьев ее вероятность значительно меньше

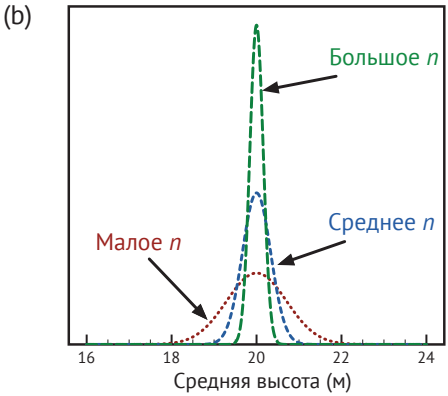


Рис. 3.3. (*Окончание*) (b) Для любого распределения генеральной совокупности существует семейство выборочных распределений, по одному для каждого размера выборки *n*. Все выборочные распределения гауссовы. Чем больше размер выборки, тем меньше стандартное отклонение выборочного распределения. По мере возрастания *n* вероятность того, что выборочное среднее будет сильно отличаться от истинного, уменьшается. Это и не удивительно, потому что если измерить высоты всех деревьев, то не будет никакой выборочной ошибки. Стандартное отклонение равно нулю. Если не удалось обмерить всего несколько деревьев, то ошибка будет очень мала

Резюме. Из-за недостаточности выборки выборочное среднее, скорее всего, отличается от истинного. Стандартная ошибка $s_{\bar{x}}$ является мерой ожидаемой выборочной ошибки.

3.2.2. Сравнение средних

Теперь посмотрим, как недостаточность выборки влияет на сравнение средних на северном и южном склоне. Очевидно, что для деревьев на южном склоне тоже существует семейство выборочных распределений. Далее мы будем предполагать, что размеры выборок и дисперсии генеральных совокупностей одинаковы для обоих склонов. Как уже отмечалось, если две выборки содержат все деревья из каждой генеральной совокупности, то мы можем просто сравнить средние и вычислить разность. Если же размеры выборок меньше, то оба выборочных средних могут сильно отличаться от истинных. Сначала вычислим разность выборочных средних: $\bar{x}_{North} - \bar{x}_{South}$. Для каждой пары выборок с северного и южного склонов мы можем сравнить разность выборочных средних с разностью истинных средних $\mu_{North} - \mu_{South}$. Таким образом, имеется только одно выборочное распределение, и ситуация оказывается такой же, как в предыдущем подразделе.

Как и прежде, выборочное распределение нормальное, а его среднее равно разности средних генеральной совокупности $\mu_{North} - \mu_{South}$. Кроме того, «правило дисперсии суммы» описывает связь стандартного отклонения этого выборочного распределения со стандартным отклонением генеральных совокупностей и размером выборки:

$$\sigma_{\bar{x}_{North}-\bar{x}_{South}} = \sigma\sqrt{\frac{2}{n}}. \tag{3.7}$$

Разность выборочных средних называется *стандартной ошибкой*[1]. Если средние и стандартное отклонение генеральных совокупностей неизвестны, то можно рассмотреть оценку:

$$s_{\bar{x}_{North}-\bar{x}_{South}} = s\sqrt{\frac{2}{n}}. \tag{3.8}$$

Вспомним первоначальный вопрос. Мы взяли по одной выборке деревьев с северного и южного склонов, обе размера n. Скорее всего, два выборочных средних различаются: $\bar{x}_{North} - \bar{x}_{South} \neq 0$. Чем обусловлена эта разница: недостаточностью выборки, притом, что истинные средние генеральных совокупностей одинаковы, или действительным различием в высотах деревьев на разных склонах? Это классическая ситуация ТОС – только вместо одиночных измерений мы имеем средние. Насколько хорошо мы способны провести различие между альтернативами? На этот вопрос можно «ответить, вычислив d' или δ Коэна для альтернатив. Для первой альтернативы $\mu_{North} - \mu_{South} = 0$, т. е. средние высоты деревьев на северном и южном склонах одинаковы. Это значим, что мы имеем распределение чистого шума. Соответствующее выборочное распределение центрировано вокруг 0, потому что различий нет. Для второй альтернативы различие присутствует, и выборочное распределение центрировано относительно $\mu_{North} - \mu_{South}$. Поскольку истинные значения неизвестны, мы используем оценки[2].

Итак, мы оценили два выборочных распределения: в одном налицо различие (сигнал и шум) со средним $\mu_{North} - \mu_{South}$, а в другом

[1] Если размеры выборок для деревьев с северного и южного склонов отличаются, то формула имеет вид:

$$\sigma_{\bar{x}_{North}-\bar{x}_{South}} = \sigma\sqrt{\frac{1}{n_{North}} + \frac{1}{n_{South}}}.$$

[2] Обычно значение s вычисляется путем объединения дисперсий для двух выборок. Один способ такого объединения мы опишем в разделе 3.4.

различия нет (только шум) и среднее равно нулю. Следовательно, мы имеем в точности такую же ситуацию, как в примере с желтой подводной лодкой, и оцениваем d' или δ Коэна выборочных распределений, которая обычно называется t, по формуле:

$$t = \frac{(\bar{x}_{North} - \bar{x}_{South}) - (0)}{s_{\bar{x}_{North} - \bar{x}_{South}}}. \quad (3.9)$$

t-значение не что иное, как d' применительно к выборочным распределениям. Как всегда в ТОС, если t велико, то сравнительно легко различить, обусловлено ли различие распределением зашумленного сигнала или чистого шума. Если t мало, то определить, действительно ли имеется различие, будет трудно[1].

Подставим оценку стандартной ошибки в уравнение t:

$$t = \frac{(\bar{x}_{North} - \bar{x}_{South}) - (0)}{s_{\bar{x}_{North} - \bar{x}_{South}}} = \frac{(\bar{x}_{North} - \bar{x}_{South})}{s\sqrt{\frac{2}{n}}} = \frac{(\bar{x}_{North} - \bar{x}_{South})}{s}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}}. \quad (3.10)$$

Разбив стандартную ошибку, являющуюся мерой выборочной ошибки, на s и n, мы видим, что t-значение равно d (оценка δ распределений генеральной совокупности), умноженному на квадратный корень из половины размера выборки.

Интерпретировать t-значение можно двумя способами. Во-первых, нас интересует, существует ли реальное различие между средними значениями. Поэтому мы вычисляем разность обоих средних и смотрим, насколько они различаются. Однако большая разность не несет никакого смысла, если велик шум, т. е. стандартное отклонение. По этой причине мы, как и в главе 2, делим разность на оценку стандартного отклонения, которая в данном случае равна оценке стандартного отклонения выборочного распределения разности средних. Оценка стандартного отклонения выборочного распределения – это стандартная ошибка:

$$s_{\bar{x}_{North} - \bar{x}_{South}} = \frac{s}{\sqrt{\frac{n}{2}}}. \quad (3.11)$$

[1] По соглашению, принятому в ТОС, t всегда интерпретируется как положительное число, если явно не оговорено противное. Если вычисленное значение отрицательно, всегда можно просто поменять порядок средних в числителе.

Таким образом, стандартная ошибка выборочного распределения средних объединяет оба источника неопределенности: стандартное отклонение генеральной совокупности и неопределенность в связи с недостаточностью выборки. Во-вторых, мы видим, что *t*-значение – это произведение оценки *d* распределения генеральной совокупности на функцию от размера выборки *n*. *t*-значение объединяет размер эффекта с размером выборки.

Резюме. Мы хотели узнать, одинаковы или нет два средних, но возникли трудности, потому что в нашем распоряжении есть только неточные оценки, – из-за недостаточности выборки. Это классическая задача на различение, только вместо одиночных измерений в ней фигурируют средние. *t*-значение, которое легко вычислить по имеющимся выборкам, не что иное, как оценка d' для этой ситуации. Но важнее, что *t*-значение – функция оценочного размера эффекта *d* и размера выборки *n*, а именно произведение *d* на квадратный корень из *n*/2.

3.2.3. Ошибки типа I и II

Недостаточность выборки может стать причиной ошибки и, как следствие, неверных выводов. Например, мы можем решить, что средние высоты деревьев на северном и южном склоне различаются (хотя на самом деле это не так), потому что различаются выборочные средние (ложная тревога). Аналогично можно решить, что они не различаются (хотя в действительности различие есть), потому что разность между выборочными средними мала (ошибочный пропуск). Следуя соглашениям, принятым в статистике, мы называем ложную тревогу ошибкой типа I, а ошибочный пропуск – ошибкой типа II. Как поступать с этими ошибками? В главе 2 мы видели, что ложные тревоги и ошибочные пропуски зависят от заданного нами порога. То же самое верно и здесь, а четыре возможных исхода показаны на рис. 3.4. Обычно исследователя интересует так называемая *нулевая гипотеза*: средние генеральных совокупностей равны. В терминах ТОС нулевая гипотеза означает, что наблюдаемое различие между выборочными средними объясняется недостаточностью выборки, а в действительности имеет место чистый шум. Альтернативная гипотеза, обозначаемая H_a или H_1, заключается в том, что средние двух генеральных совокупностей различны. В терминах ТОС это означает, что наблюдаемое различие между выборочными средними объясняется распределением зашумленного сигнала. На рис. 3.4 эта ситуация обозначена фразой «H_0 неверна».

Как и в примере с желтой подводной лодкой, большое *t* означает, что различить средние генеральных совокупностей будет просто, а малое значение – что различение затруднено и может приводить к ошибочным выводам. Теперь легко принять решение о нулевой гипотезе. Мы вычисляем *t*, а затем применяем порог. Если вычисленное значение *t* больше порога, то это расценивается как свидетельство в пользу того, что оценочная разность средних не объясняется распределением чистого шума: между двумя средними действительно имеется различие. Если вычисленное значение *t* меньше порога, то нет уверенности, что средние различны. Может быть, различны, а может быть, и нет. Никакого определенного вывода сделать нельзя.

Рис. 3.4. Задача статистики – делать заключения о гипотезе. Как и в главах 1 и 2, возможно четыре исхода. (1) Нулевая гипотеза неверна, и мы пришли к выводу, что средние различны (Правильное подтверждение). (2) Нулевая гипотеза неверна, но у нас недостаточно фактов, свидетельствующих против нее (Ошибочный пропуск, или ошибка типа II). (3) Нулевая гипотеза верна, и мы пришли к выводу, что средние различны (Ложная тревога, или ошибка типа I). (4) Нулевая гипотеза верна, и у нас недостаточно фактов, свидетельствующих против нее (Правильный пропуск)

На практике в различных областях используются разные пороги, отражающие наиболее подходящие уровни правильного подтверждения или ложной тревоги. Например, в физике часто применяется критерий пяти сигм, согласно которому различие между средними считается экспериментально доказанным, если $t > 5$. По сравнению с другими областями это очень высокое значение; отчасти оно отражает тот факт, что у физиков часто имеется возможность (и ресурсы) существенно уменьшить σ и s за счет усовершенствования техники измерений. В физике элементарных частиц большой адронный коллайдер порождает выборки триллионного размера. В таких областях, как медицина, психология, нейронауки и биология, обычно приме-

няется порог, который в первом приближении следует «правилу двух сигм». Выбор менее строгого порога в какой-то мере связан с условиями научных исследований в этих областях. В некоторых представляющих интерес вопросах шума в принципе не избежать, а различия между генеральными совокупностями малы. При этом стоимость получения одного образца для медицинской или биологической выборки часто гораздо выше, чем в физике, а в некоторых ситуациях (например, при исследовании пациентов с редкими заболеваниями) набрать выборки большого размера вообще невозможно.

ТОС также говорит нам, что какой бы порог ни выбрать, имеет место компромисс между правильными подтверждениями и ложными тревогами, и критерии пяти или двух сигм не исключение. При прочих равных условиях критерий пяти сигм будет давать меньше правильных подтверждений, чем критерий двух сигм. Но зато он будет давать и меньше ложных тревог.

Вместо того чтобы задавать порог в терминах стандартного отклонения σ, во многих областях (включая медицину, психологию, нейронауки и биологию) исследователь хочет, чтобы частота ошибок типа I была меньше заранее заданного значения, например 0.05. Понятно, почему возникает желание ограничить ошибку этого вида: она побуждает человека верить в существование эффекта, которого на самом деле нет. Например, можно прийти к выводу, что лечение помогает пациенту, тогда как в действительности оно неэффективно и следует попробовать другое лекарство. Такие ошибки могут стать причиной смерти. С философской точки зрения ученые стоят на позициях скептицизма и по умолчанию считают, что различий нет: лечение не помогает больному, вмешательство не помогает образованию, мужчины и женщины имеют схожие признаки. Ученый отступает от скептической позиции, только если имеется достаточно свидетельств в пользу ее необоснованности.

Резюме. Ошибка типа I возникает, когда средние равны, т. е. нулевая гипотеза верна, но мы решили, что они разнятся. Ошибка типа I – то же самое, что ложная тревога в контексте ТОС. Чтобы принять решение относительно нулевой гипотезы, мы вычисляем величину t, эквивалентную различимости в терминах ТОС, а затем применяем порог.

3.2.4. Ошибка типа I: *p*-значение связано с порогом

В этом разделе мы покажем, что порог определяет частоту ошибок типа I. Рассмотрим, как выглядит выборочное распределение разности выборочных средних, когда нулевая гипотеза H_0 верна, т. е. $\mu_{North} - \mu_{South} = 0$.

Распределение центрировано относительно нуля со стандартной ошибкой, которую мы оценили на основе данных. Предположим, что задан порог $t = 2.0$, который часто называют критическим значением (critical value – cv) и обозначают $t_{cv} = 2.0$. Если для наших данных t-значение больше $t_{cv} = 2.0$, то мы заключаем, что выборочные средние различны, даже если это не так, т. е совершаем ошибку типа I. Вероятность такого t-значения равна площади под кривой в области правее $t_{cv} = 2.0$ (см. рис. 3.5a). Такой тест называется «односторонним t-критерием». Вычисление площади этой области (в предположении большого размера выборки) дает 0.0228. Следовательно, при использовании порога $t_{cv} = 2.0$ мы допустим ошибку типа I с вероятностью всего 0.0228. Если поступать так, как принято в физике, т. е. использовать критерий пяти сигм, то $t_{cv} = 5$, и вероятность ошибки типа I составит 0.000000287.

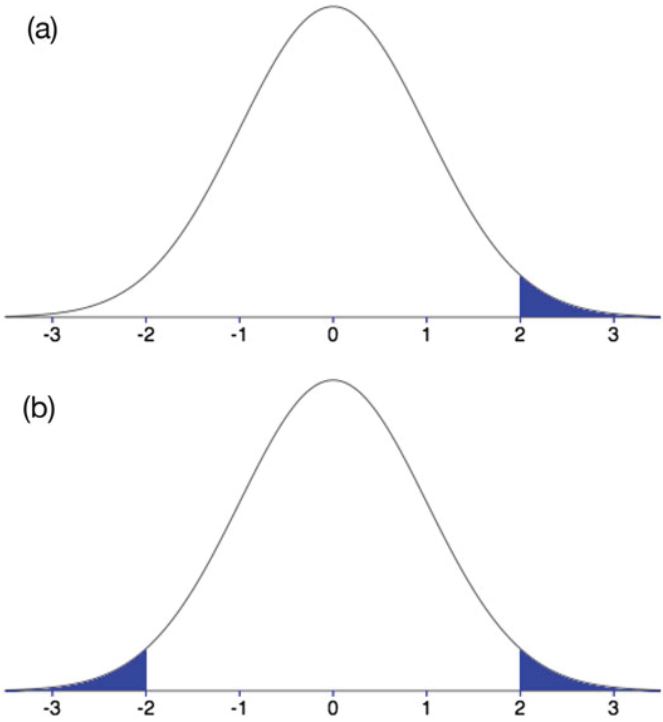


Рис. 3.5. Связь между критическим значением порога и частотой ошибок типа I. Кривая показывает распределение чистого шума, т. е. случай, когда нулевая гипотеза верна. Распределение центрировано относительно нуля, а дисперсия оценена по данным. (a) При критическом значении $t_{cv} = 2.0$ частота ошибок типа I равна площади под кривой в области правее t_{cv}. Этот тест называется односторонним t-критерием. (b) При критическом значении $t_{cv} = \pm 2.0$ частота ошибок типа I равна площади под кривой в области правее 2.0 и левее –2.0

Этот подход обладает большой гибкостью. Например, мы можем заподозрить, что деревья на северном и южном склоне имеют разную высоту, но не знаем, где они выше. В таком случае можно использовать порог $t_{cv} = \pm 2.0$, где t-значение, более экстремальное (расположенное дальше от нуля), чем 2.0, будет считаться свидетельством в пользу того, что средние генеральных совокупностей различны. При таком подходе частота ошибок типа I составила бы 0.0456, т. е. в два раза больше, чем в случае одностороннего критерия (см. рис. 3.5b). Этот тест называется двусторонним t-критерием.

В примере выше мы задали порог и вычислили для него частоты ошибок типа I. Но обычно в статистике происходит наоборот. Фиксируется частота ошибок типа I, и вычисляется соответствующее ему t-значение порога. Например, если принять частоту ошибок типа I, равную 5 %, то соответствующее t-значение порога будет равно $t_{cv} = \pm 1.96$ для двустороннего t-критерия, если n велико[1]. Вместо того чтобы задавать некоторое t-значение в качестве порога, исследователь вычисляет площадь под кривой за пределами t-значения, вычисленного по данным. Эта площадь называется p-значением (см. рис. 3.6). Таким образом, мы сначала вычисляем t-значение на основе данных, а затем p-значение. p-значение говорит, насколько вероятно при условии истинности нулевой гипотезы получить наше или даже большее t-значение. Если p-значение меньше 0.05, то эффект называется значимым. Следовательно, для контроля частоты ошибок типа I нужно лишь потребовать, чтобы вычисленное p-значение было меньше желаемой частоты ошибок.

Как уже отмечалось, в медицине, психологии, нейронауках и биологии часто задают желаемую частоту равной 0.05. Для большого размера выборки и в ситуациях, когда рассматриваются как положительные, так и отрицательные t-значения (двусторонний t-критерий), значение $p = 0.05$ соответствует $t = \pm 1.96$. Стало быть, задание частоты ошибок типа I, равной 0.05, соответствует порогу $t_{cv} = \pm 1.96$. Именно в силу этого соотношения в указанных областях применяют правило двух (приближенно) сигм. t-значение можно вычислить вручную, но для вычисления p-значения нужна статистическая программа.

Резюме. Если t-значение больше некоторой величины (зависящей от размера выборки n), то мы заключаем, что имеется значимый эффект.

---

[1] Для выборок небольшого размера значение t_{cv} больше, потому что выборочное распределение не совсем нормальное. Статистические программы, вычисляющие p-значение, автоматически вносят корректировки, учитывающие отклонение выборочных распределений от нормальных.

3.2.5. Ошибка типа II: подтверждения, пропуски

В общем случае TOC говорит, что для данного d' задание порога не только определяет частоту ошибок типа I, но и устанавливает частоты правильных подтверждений, ошибочных пропусков и правильных пропусков. Действительно, легко видеть, что использование порога, которому соответствует частота ошибок типа I 0.05, также определяет частоту правильных пропусков (когда мы заключаем, что свидетельств в пользу эффекта недостаточно, и действительно эффект отсутствует), равную 1.0 – 0.05 = 0.95. Как показывает синяя область на рис. 3.7, для данного порога площадь левого и правого хвоста под этой кривой альтернативного выборочного распределения соответствует вероятности взять выборки, порождающие правильное подтверждение (порог превышен, и делается вывод о достаточном свидетельстве в пользу различия средних генеральных совокупностей) и ошибку типа II (порог превышен, и не делается вывод о достаточном свидетельстве в пользу различия средних генеральных совокупностей).

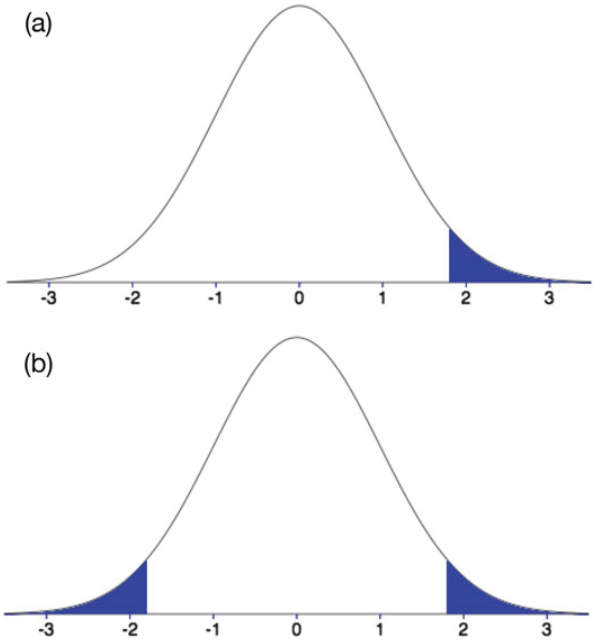


Рис. 3.6. Связь между статистикой *t*-критерия и *p*-значениями. (a) Для одностороннего критерия с $t = 1.8$ *p*-значение – это площадь под кривой правее 1.8, оно равно 0.0359. (b) Для одностороннего критерия с $t = 1.8$ мы вычисляем площадь обоих хвостов – правее +1.8 и левее –1.8. В этом случае *p*-значение равно 0.0718

Поэтому создается впечатление, будто вычислить частоту ошибок типа II было бы столь же просто. Но это не так. При вычислении частоты ошибок типа I мы знаем, что выборочное распределение, соответствующее нулевой гипотезе, центрировано относительно одного значения, а именно 0. Следовательно, существует только одна нулевая гипотеза. Однако альтернативных гипотез бесконечно много (см. следствие 2e). Но, быть может, нас интересуют только значительные различия между средними высотами на северном и южном склоне, когда северные деревья выше южных не менее чем на 1.2 м. Коли так, то мы знаем минимальное расстояние между распределениями генеральных совокупностей и можем спросить, каким должен быть размер выборки *n*, чтобы значимый результат достигался как минимум в 80 % случаев.

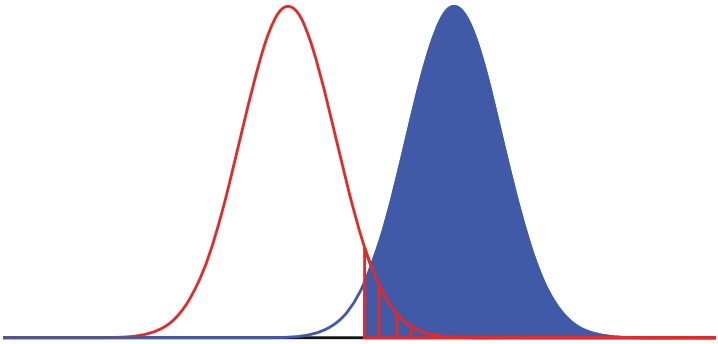


Рис. 3.7. Выборочные распределения и вычисление частоты правильных подтверждений. Порог соответствует нижней границе красной заштрихованной области и нижней границе синей сплошной области. Любое *t*-значение, оказывающееся выше этого порога, приводит к решению отвергнуть H_0 и заключить, что средние генеральных совокупностей различны. Площадь синей области – вероятность того, что выборочное распределение H_a дает *t*-значение из этого диапазона. Частота ошибок типа I – площадь заштрихованной красным области под красной кривой

Частота ошибок типа II играет важную роль с точки зрения мощности. Мощностью называется вероятность получить значимый результат, когда эффект действительно имеет место, т. е. когда верна альтернативная гипотеза. Мощность – это просто другой термин для частоты правильных подтверждений. Частота правильных подтверждений равна 1 минус частота ошибок типа II. Мощность окажется в центре нашего внимания в части III и подробнее объясняется в главе 7.

3.3. Резюме

Все сказанное выше – фундаментальные вещи, важные для понимания статистики. Поэтому мы еще раз пройдемся по основным шагам и отметим наиболее важные моменты. Даже если вы не читали предыдущих разделов, главные идеи будут понятны из этого краткого изложения.

Нас интересовало, верно ли, что *средняя* высота дубов на северном склоне Альп такая же, как на южном. На этот вопрос ответить просто. Нужно лишь обмерить все деревья, вычислить оба средних и посмотреть, равны они или нет. Если какие-то деревья пропустить, то получатся оценки, которые, скорее всего, отличаются от истинных средних. Чем меньше деревьев обмерено, тем больше *выборочная ошибка*, т. е. тем выше вероятность, что выборочные средние значительно отличаются от истинных. Мы показали, что эту выборочную ошибку можно количественно выразить с помощью стандартной ошибки $s_{\bar{x}}$, зависящей от стандартного отклонения генеральной совокупности σ и размера выборки n. Если σ мало, то нужно выбрать всего несколько деревьев, чтобы получить хорошую оценку среднего. Например, если $\sigma = 0$, то достаточно выбрать по одному дереву из каждой популяции, потому что все деревья имеют одинаковую высоту. Если σ велико, то для получения хорошей оценки среднего нужно взять много деревьев.

Теперь сформируем выборки деревьев с северного и южного склона Альп размера n, гораздо меньшего, чем общее число деревьев. Вычислим среднюю высоту в обеих выборках. Из-за недостаточности выборки выборочные средние почти наверняка будут различны. Однако мы не знаем, объясняется ли наблюдаемое различие недостаточностью выборки, или же средние генеральных совокупностей действительно различны. Если истинные средние одинаковы, но на основе различия выборочных средних мы пришли к выводу, что они различны, то мы допустили ошибку типа I (рис. 3.4). Ученые обычно стараются избегать ошибок типа I, потому что по умолчанию считают, что эффекта нет, если только данные явно не указывают на обратное. Никакой процесс принятия решений не может вовсе избежать ошибок типа I, но мы в состоянии контролировать частоту таких ошибок. В этом отношении важную роль играет t-значение, которое легко вычислить вручную:

$$t = \frac{(\bar{x}_{North} - \bar{x}_{South})}{s}\sqrt{\frac{n}{2}} = d\sqrt{\frac{n}{2}}. \qquad (3.12)$$

Мы вычисляем выборочные средние x_{North} и x_{South} по обмерам n деревьев x_i, оцениваем стандартное отклонение s высот деревьев (серый квадратик) и умножаем на функцию (квадратный корень) от половины размера выборки $n/2$. Правая часть показывает, что t-значение не что иное, как оценка размера эффекта d, умноженная на функцию от размера выборки. t-значение говорит, насколько легко установить, является ли различие выборочных средних следствием реального различия средних генеральной совокупности. Ситуация в точности такая же, как в главе 2. t-значение – это просто d', только вместо деления на стандартное отклонение мы делим на стандартную ошибку, являющуюся мерой выборочной ошибки, которая принимает во внимание шум, дисперсию генеральной совокупности и недостаточность выборки. Большое t-значение означает, что различить средние легко, а малое – что принять решение трудно. Отметим, что большое t-значение может иметь место, потому что что велик размер эффекта d, потому что велико n или по обеим причинам сразу.

Предположим, что никакого эффекта нет, т. е. средняя высота северных и южных деревьев одинакова ($\delta = 0$). Тогда p-значение говорит, насколько вероятно, что случайная выборка даст t-значение, не меньшее только что вычисленного. Таким образом, если нас устраивает 5%-ная частота ошибок типа I и p-значение меньше 0.05, то разность средних называется значимой.

p-значение полностью определено t-значением и вычисляется статистическими программами. Важнее то, что t-значение объединяет оценку размера эффекта d с размером выборки ($\sqrt{n}/2$), и именно поэтому t-значение, а вместе с ним и p-значение смешивает в одну кучу размер эффекта и размер выборки, а следовательно, дает только частичную информацию! Это замечание поможет понять некоторые следствия, о которых мы расскажем после разбора следующего примера.

3.4. Пример

Вычислить p-значение просто, как показано в следующем примере. А вот понять следствия, вытекающие из t-критерия, сложнее.

Предположим, что мы измерили высоты пяти деревьев на северном склоне и пяти деревьев на южном. Данные представлены в первом столбце на рис. 3.8. Там же приведены вычисления для двустороннего t-критерия. Для заданных размеров выборок и вычисленного t-значения наша статистическая программа сообщает, что соответствующее p-значение равно 0.045. Поскольку это p-значение меньше

0.05, мы заключаем, что данные свидетельствуют о значимом различии между средними высотами северных и южных деревьев[1].

Результаты подобных тестов часто сводятся в таблицу типа табл. 3.1. *p*-значение находится в столбце «Знач. (двусторонний)». В таблице также упомянуто число степеней свободы (*df*). Оно важно для вычисления *p*-значения, потому что форма выборочного распределения немного отличается от нормальной для выборок малого размера. Величину *df* можно вычислить по размеру выборки, и наоборот. В случае *t*-критерия $df = n_1 + n_2 - 2 = 5 + 5 - 2 = 8$.

Как уже отмечалось, значимость мало что говорит о полученных результатах. Важно включать в отчет размер эффекта. Коэн предложил рекомендации для величин эффектов в случае *t*-критерия, они приведены в табл. 3.2.

Что следует запомнить

- Поскольку *p*-значение определяется *t*-значением, оно смешивает в одну кучу размер эффекта (*d*) и размер выборки (*n*). Изначально предполагалось, что *t*-критерий даст инструмент для понимания того, в какой степени значимый результат является следствием случайной выборки при заданном размере эффекта *d*. В настоящее время *p*-значение часто ошибочно используется как показатель размера эффекта, хотя оно никогда не задумывалось для этой цели и такое применение попросту неверно!
- Частичная информация: к правильным выводам можно прийти, только принимая во внимание как оценку размера эффекта в генеральной совокупности *d*, так и размер выборки *n*. Поэтому важно включать в отчет оба значения – как основу для умозаключений и чтобы понимать, вызван ли значимый результат оценочным размером эффекта *d*, размером выборки или тем и другим одновременно.

[1] Вместо этого можно было бы найти критическое значение порога, $t_{cv} = \pm 2.306$, и заметить, что *t* отстоит от нуля дальше, чем это значение.

Север	$(x_{i,North} - \bar{x}_{North})^2$
20.80	$(20.80 - 18.00)^2 = 7.840$
17.81	0.036
17.92	0.006
18.30	0.090
15.17	8.009
$\bar{x}_{North} = \sum_{i=1}^{n} \frac{x_{i,North}}{n} = 18.00$	$s^2_{North} = \frac{\sum_{i=1}^{n} (x_{i,North} - \bar{x}_{North})^2}{n-1} = 3.995$

Юг	$(x_{i,South} - \bar{x}_{South})^2$
16.91	3.648
15.28	0.078
13.70	1.690
16.81	3.276
12.30	7.290
$\bar{x}_{South} = \sum_{i=1}^{n} \frac{x_{i,South}}{n} = 15.00$	$s^2_{South} = \frac{\sum_{i=1}^{n} (x_{i,South} - \bar{x}_{South})^2}{n-1} = 3.996$

$$df = (5-1) + (5-1) = 4 + 4 = 8$$

$$s_p^2 = \frac{s^2_{North}(n-1) + s^2_{South}(n-1)}{(n-1)+(n-1)} = \frac{3.995(4) + 3.996(4)}{4+4} = 3.996$$

$$s_{\bar{x}_{North} - \bar{x}_{South}} = \sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{n}\right)} = \sqrt{3.996\left(\frac{1}{5} + \frac{1}{5}\right)} = 1.264$$

$$t = \frac{\bar{x}_{North} - \bar{x}_{South}}{s_{\bar{x}_{North} - \bar{x}_{South}}} = \frac{18.00 - 15.00}{1.264} = 2.373$$

Рис. 3.8. Вычисления *t*-критерия для независимых выборок начинаются с вычисления средних по каждому столбцу ($\bar{x}_{North}$ и $\bar{x}_{South}$, показаны в последних строках первого столбца). Зная их, мы можем вычислить дисперсии (s^2_{North} и s^2_{South} в последних строках второго столбца). Число степеней свободы (*df*) для каждого столбца вычисляется как число примеров в столбце минус единица. Затем вычисляется объединенная дисперсия (s^2_p) как взвешенная сумма двух дисперсий (в роли веса выступает число степеней свободы). Далее объединенная дисперсия подставляется в формулу стандартной ошибки $s_{\bar{x}_{North} - \bar{x}_{South}}$, и результат используется в знаменателе нашей формулы для *t*. В числителе стоит просто разность между средними, вычисленными на первом шаге

Таблица 3.1. Результат типичного пакета статистических программ

	t	df	Знач. (двусторонний)	d Коэна
Высота дерева	2.373	8	0.045	1.5 (большой эффект)

Столбцы *t*, *df* и *Знач. (двусторонний)* содержат *t*-значение, соответствующее ему число степеней свободы и *p*-значение. Значение *df* здесь равно сумме df_N и df_S (т. е. 4 + 4 = 8).

Таблица 3.2. Рекомендации Коэна по размеру эффекта *d*

	Малый	Средний	Большой
Размер эффекта	0.2	0.5	0.8

3.5. Следствия, комментарии и парадоксы

Для описанных ниже следствий особенно важна формула (3.12), потому что она говорит, что t-значение, а стало быть, и p-значение определяются оценочным d и размером выборки n.

Следствия 1. Размер выборки

Следствие 1a. Согласно формуле (3.12), если оценочное значение $d \neq 0$, то всегда найдется n, для которого t-критерий будет значимым. Поэтому даже при очень малом размере эффекта может получиться значимый результат, если размер выборки достаточно велик. Поэтому к значимым результатам приводит не только большой размер эффекта, как могло бы показаться, но и любой отличный от нуля размер эффекта при условии, что n достаточно велико[1].

Следствие 1b. Если оценочное значение $d \neq 0$ (и $d < 4.31$), то существует размер выборки $n < m$ такой, что t-критерий не является значимым для n, но является значимым для m[2]. Это утверждение может показаться парадоксальным, если прочитать его следующим образом: для n эффекта не существует, а для m существует. Однако такое прочтение некорректно. Мы можем только заключить, что для m имеется достаточно свидетельств в пользу значимого результата, а для n таких свидетельств недостаточно. Из нулевого результата (когда мы не отвергаем нулевую гипотезу) нельзя сделать никаких выводов (см. следствие 3). В части III мы увидим, что этот кажущийся парадокс указывает на ключевую проблему проверки гипотез.

Следствие 1c. Провокационный вопрос: «Разве не всегда существует различие между двумя условиями, пусть совсем крохотное?» Кажется, что, за исключением немногих случаев, различие между средними двух генеральных совокупностей $\mu_1 - \mu_2$ никогда не обращается в нуль. Сами посудите – насколько вероятно, что обе средние высоты деревьев – на северном и южном склоне – в точности равны 20.2567891119 м? Но раз так, то мы всегда сможем найти такой размер выборки n, при котором результаты эксперимента будут значимыми. А зачем тогда нужны эксперименты?

---

[1] Эту ситуацию можно также описать следующим образом. Если действительно имеет место эффект $d \neq 0$, например средние высоты деревьев различаются, то можно найти такой размер выборки n, при котором значимый результат получается почти во всех случаях (или с очень высокой вероятностью).

[2] Если $d > 4.31$ то статистику вычислять не нужно, потому что разность велика. В этом случае даже $n = 2$ приводит к значимому результату.

Следствия 2. Размер эффекта

Следствие 2a. Как уже отмечалось, *p*-значение не является способом измерения размера эффекта в генеральной совокупности δ, и для любого $d \neq 0$ существует такое n, при котором результат является значимым. Таким образом, даже малые эффекты могут быть значимыми. Согласно некоторым исследованиям ежедневное употребление рыбьего жира может значимо продлить жизнь. Но, быть может, всего на две минуты. Вам это важно?

Следствие 2b. Само по себе *p*-значение ничего не говорит о размере эффекта. Например, с увеличением размера выборки (при прочих равных условиях) *p*-значение уменьшается, потому что уменьшается дисперсия выборочного распределения (см. рис. 3.3). Таким образом, если размер эффекта *d* остается неизменным, то *p*-значение изменяется в зависимости от размера выборки.

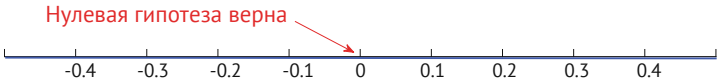


Рис. 3.9. В терминах размера эффекта нулевая гипотеза представлена ровно одной, нулевой, точкой. Все остальные точки, коих бесконечно много, относятся к гипотезе H_1

Следствие 2c. *p*-значения двух экспериментов A и B могут совпасть. Но из этого факта нельзя сделать никаких выводов. Например, может случиться так, что в эксперименте A был большой размер эффекта *d* и малый размер выборки, а в эксперименте B наоборот. Следовательно, при разных размерах выборки сравнивать *p*-значения двух экспериментов нельзя. Аналогично если в эксперименте A *p*-значение меньше, чем в B, то это не значит, что размер эффекта больше. Просто мог быть больше размер выборки.

Следствие 2d. Если в исследовании с небольшим размером выборки получилось малое *p*-значение, значит, оценочный размер эффекта меньше, чем в исследовании с бóльшим размером выборки и таким же *p*-значением.

Следствие 2e. Еще раз повторим и наглядно проиллюстрируем базовую ситуацию. Худший случай – когда нулевая гипотеза верна, т. е. средние одинаковы, а мы заключаем, что они различны, т. е. совершаем ошибку типа I. В этом случае $\mu_1 - \mu_2 = 0$. Если нулевая гипотеза неверна, то $\mu_1 - \mu_2$ отлично от 0 и, в принципе, может быть любым значением от $-\infty$ до $+\infty$. Все эти значения являются частью альтер-

нативной гипотезы, согласно которой высота деревьев на северном и южном склоне различается. Таким образом, беспокоясь по поводу ошибки типа I и нулевой гипотезы, мы высказываем опасения относительно одной-единственной точки среди бесконечного множества других точек (см. рис. 3.9).

Следствия 3. Нулевые результаты

Следствие 3a. Отсутствие доказательства – еще не доказательство отсутствия: нельзя сделать вывод об отсутствии эффекта в эксперименте ($d = 0$), если не было значимого результата. Незначимое p-значение говорит, что либо различия нет, либо оно есть, но слишком мало для достижения значимости при заданном размере выборки n.

Следствие 3b. Разность значимостей не то же самое, что значимая разность. Рассмотрим исследование по измерению воздействия крема, содержащего алоэ вера, на кожную экзему. Пациенты с экземой случайным образом распределены по двум группам: одна получает крем с содержанием алоэ вера, другая – плацебо. По истечении четырех недель повторно измеряется размер экземы. Имело место значимое уменьшение в группе, получавшей алоэ вера, но не в контрольной группе (рис. 3.10). Возникает соблазн заключить, что алоэ вера излечивает экзему. Однако в контрольной группе тоже имело место уменьшение, только не такое значительное (возможно, объясняющееся самоизлечением). На самом деле после вычисления различия в уменьшении экземы для каждого участника исследования в обеих группах и построения двустороннего t-критерия между группами различие оказывается незначимым.

Можно возразить, что при большем размере выборки различие между двумя группами могло бы оказаться значимым. Быть может, и так. Однако мог бы также оказаться значимым и эффект плацебо. И какой вывод мы должны сделать? Наперекор интуиции эта ситуация не составляет проблемы, потому что мы можем спросить, верно ли, что алоэ вера дает более сильный эффект, чем плацебо (и таким образом свести на нет самоизлечение).

Этот пример показывает, что часто не имеет особого смысла сравнивать утверждения типа «эффект наблюдался при условии A, но не при условии B». Такие умозаключения встречаются в науке повсеместно, но относиться к ним следует с большой осторожностью (см. также главу 7). Классический пример – как в случае с алоэ вера, сравнить последствия вмешательства, при котором ожидаются значимые результаты, с контрольным экспериментом, в котором ожидается получение нулевого результата.

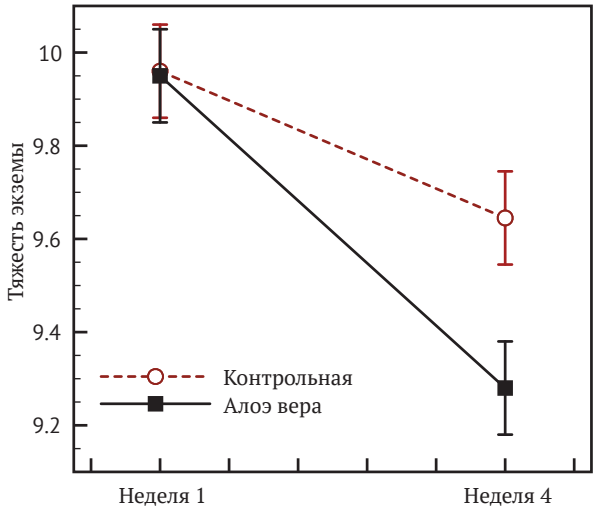


Рис. 3.10. В исследовании воздействия алоэ вера на экзему экспериментальная группа получала крем, содержащий алоэ вера, а контрольная группа – плацебо. Тяжесть экземы измерялась в начале эксперимента и спустя четыре недели. На графике показана средняя тяжесть экземы для каждой группы при каждом измерении. «Усы» обозначают стандартную ошибку каждого среднего. Тяжесть экземы значимо уменьшилась в экспериментальной группе, но не в контрольной. Можем ли мы заключить, что алоэ вера излечивает экзему? Не можем, потому что тяжесть заболевания уменьшилась и в контрольной группе, возможно, вследствие самоизлечения. На самом деле никакого значимого эффекта не было выявлено, когда улучшения в обеих группах сравнили с помощью двустороннего *t*-критерия. Разность значимостей не то же самое, что значимая разность. На графике показаны средние значения и соответствующие стандартные ошибки (см. врезку «Некоторые определения»)

Следствия 4. Истина, шум и изменчивость

Следствие 4а. Почему вообще мы вычисляем статистику? Часто неявно предполагается, что статистика «очищает» от шумов, неизбежных в сложных системах. В примере с подводной лодкой на результат измерения оказывали влияние изменения в толще воды, например рыбы и водоросли, или в самом устройстве, подверженном случайным флуктуациям. Такого рода шум называется шумом измерений. Все источники шума искажают истинный сигнал как при наличии скалы, так и при ее отсутствии. Ситуацию можно описать следующей формулой:

$$x_j = \mu + \varepsilon_j,$$

где x_j – результат j-го измерения, μ – истинный сигнал, а ε_j – шум, зависящий от испытания. Обычно предполагается, что ε_j имеет нор-

мальное распределение с нулевым средним. Таким образом, класть в основу решения одно испытание – неудачная мысль. Как уже отмечалось, усреднение по многим измерениям может устранить шум. Поэтому лучше сравнивать средние значения, а не одиночные измерения. Как мы видели, чем больше n, тем точнее измерение среднего.

Модель такого вида пригодна во многих областях, в частности в физике. Однако в биологии, медицине и других науках ситуация зачастую бывает совершенно иной. Например, мы могли бы определить силу головной боли до и после приема болеутоляющего. И обнаружить, что лекарство уменьшает боль *в среднем*. Но, как почти всегда бывает с лекарствами, встречаются люди, на которых лекарство вообще не действует. Кроме того, на одних оно действует лучше, на других хуже; у кого-то головная боль почти проходит, а на кого-то лекарство оказывает лишь слабый (или вовсе противоположный) эффект.

Зависящие от человека эффекты можно описать следующей формулой:

$$x_{ij} = \mu + v_i + \varepsilon_{ij},$$

где x_{ij} – одно измерение, например пациент i принял болеутоляющее в день j, μ – среднее значение по всей генеральной совокупности, например в какой степени лекарство уменьшает головную боль в среднем, а v_i – чувствительность пациента i к данному болеутоляющему. Как уже было сказано, одним людям лекарство всегда помогает, на других не оказывает никакого действия, а у кого-то боль может даже усилиться. Таким образом, v_i определяет, насколько один человек отличается от других – и от среднего μ. ε_{ij} – шум измерений, он отражает, например, различие в воздействии лекарства на одного и то же человека в разные дни. В некотором смысле ε_{ij} улавливает несистематическую изменчивость, а v_i – систематическую. Рассмотрим еще один пример. У одного человека кровяное давление может быть выше, чем у другого, и это различие отражается в групповой изменчивости v_i. В то же время кровяное давление у одного и того же человека может сильно отличаться в соседние минуты, и это отражается членом ε_{ij}, характеризующим персональную изменчивость.

Во многих экспериментах разделить v_i и ε_{ij} нелегко. Оба члена вносят вклад в оценочное стандартное отклонение распределения генеральной совокупности s. С точки зрения математики неважно, имеет ли место сильная групповая изменчивость или сильный шум измерения. Но для интерпретации результатов статистического анализа это различие принципиально важно. Предположим, что

болеутоляющее оказывает сильный благоприятный эффект на половину генеральной совокупности и слабое негативное воздействие на другую половину. В среднем эффект лекарства положительный, и этот эффект может оказаться значимым. Но важно, что, несмотря на в среднем положительный эффект, для конкретного человека это может быть не так. Для половины генеральной совокупности эффект негативный, поэтому использовать такое болеутоляющее не стоит. Следовательно, когда v_i отлично от нуля, значимые результаты не позволяют делать заключения на уровне индивидуума. Исследование может показать, что морковь в среднем хороша для зрения. Но так ли это лично для вас, неясно. Морковь может даже ухудшить ваше зрение, хотя другим людям помогает. Эти соображения не означают, что все подобные исследования бессмысленны, они просто указывают на ограничения в случае, когда $v_i \neq 0$ для некоторого i. Для международных исследований уровня кровяного давления средние значения вполне удовлетворительны. Но сравнивать себя с такой большой группой обычно неразумно, что бы ни измерялось. Мало того что такая выборка неоднородна и зависит от региона, так она еще и включает людей разных возрастов. Индекс массы тела 27 может указывать на проблему для детей младше 5 лет, но необязательно для лиц старше 70 лет. Следовательно, осмысленность сравнения средних очень сильно зависит от предмета исследования. Это вопрос интерпретации, а не вычисления статистики.

Следствие 4b. У рассмотренных выше соображений имеются и философские следствия. Обычно мы предполагаем, что некое явление либо есть, либо нет. Сила притяжения действует либо на всю материю во Вселенной, либо не действует. Кислород либо необходим человеку, либо нет. Все эти факты имеют место для каждого отдельного субъекта, т. е. для каждого элемента Вселенной, для каждого человека и т. д. Если некоторый факт был установлен с применением методов, включающих статистику, то этот вывод необязательно обоснован, когда v_i отлично от 0, потому что результаты справедливы лишь в среднем, а не для каждого конкретного субъекта.

Следствие 4c. Проблема изменчивости и шума становится еще серьезнее, если исследование имеет дело с неоднородной выборкой, систематически изменяющейся по признаку, который явно не учтен. Например, из частоты посещения врача можно сделать вывод, что студенты небольшого роста болеют чаще, чем более высокие. Однако этот факт не имеет ничего общего с ростом. Просто студентки в среднем ниже студентов-мужчин и посещают гинеколога

чаще, чем мужчины – уролога. Однако женщины посещают гинеколога в основном в профилактических целях, а вовсе не болеют чаще мужчин. Поскольку студенты вообще ходят по врачам очень редко, статистический вес визитов к гинекологу оказывается велик. Понятно, как ошибки интерпретации могут возникнуть в таких простых примерах. Но в более сложных случаях обнаружить их сложнее. И к слову, стоит задаться вопросом, так ли правильно делать выводы о частоте заболеваний на основе данных о посещении врачей.

Следствие 4d. К проблеме связи между изменчивостью и шумом можно подойти и с другой стороны. Планируя эксперимент, необходимо указывать, кого в него включать. Для большей репрезентативности хорошо делать выборку из всей генеральной совокупности, например из всего населения одной страны или всего мира. Однако при такой процедуре генеральная совокупность рискует оказаться неоднородной, и делать какие-то выводы будет труднее. Нужно ли включать космонавтов или пациентов, находящихся в коме? А как насчет больных? Большой доли людей с повышенным кровяным давлением? Чем больше частей генеральной совокупности вы исключите, тем менее репрезентативной будет выборка. И в конечном итоге может оказаться, что в ней нет никого, кроме вас.

Следствие 4e. И последнее. Эффект часто зависит от дозировки, т. е. разные люди по-разному реагируют на разные дозы. Для кого-то болеутоляющее может давать положительный эффект в малой дозе, но принесет вред при ее увеличении. Следовательно, имеет место не только систематическая групповая изменчивость, но и систематическая персональная изменчивость в дополнение к несистематическому шуму ε_{ij}. Во многих экспериментах имеется много источников, т. е. эффект зависит от дозировки, индивидуальных различий и шума, и это существенно ограничивает возможность делать выводы. Ниже мы увидим, что эффекты, зависящие от дозировки, лучше описывать с помощью корреляции (глава 8), а не t-критериев.

Следствия 5.

Следствие 5a. Парадокс статистики и опасности когортных исследований.

Для больших размеров эффекта, какие часто встречаются, например, в физике, вычислять статистику зачастую не нужно. Аналогично гипотеза о том, что слоны в среднем больше муравьев, тоже не нуждается в статистике, потому что любой живущий на земле слон больше любого муравья, так что δ очень велико. Изначально идея статистики

состояла в том, чтобы установить, действительно ли «немного зашумленный эффект» существует, и определить размер выборки n, необходимый для демонстрации существования эффекта. Можно сказать, что статистика была разработана для эффектов и выборок среднего размера. В прошлом обычно было невозможно получить значимые результаты для эффектов небольшого размера, потому что данных было мало, а обработка больших выборок оказывалась слишком громоздкой. Поэтому n, как правило, было мало, и только эксперименты с эффектами большого размера давали значимые результаты. С тех пор положение дел полностью изменилось, потому что собирать данные стало гораздо дешевле, и теперь мы можем комбинировать и обрабатывать миллионы примеров, как, например, в генетике. Поэтому сегодня статистика широко используется не только для средних, но и для очень малых размеров эффекта. Однако такое развитие не свободно от опасностей. Во-первых, не следует путать большой размер выборки с большим размером эффекта (следствие 2а). Во-вторых, зачастую очень трудно прийти к каким-то выводам, особенно в так называемых когортных исследованиях. Например, в когортных исследованиях пациенты сравниваются с контрольной группой или вегетарианцы с мясоедами. Обе группы определяются присвоенной меткой.

Рассмотрим пример. Начиная с 1948 года, в небольшом городке Фрейминге в штате Массачусетс было измерено кровяное давление у 5209 участников. На рис. 3.11 по оси x отложен возраст участников, а по оси y систолическое давление. Данные разбиты на группы по уровню образования. Во-первых, мы видим отчетливый эффект возраста. Во-вторых, чем выше образование, тем ниже кровяное давление. Применяя статистические методы, описанные в главе 6, можно заключить, что эффект образования значимый. Означает ли это, что чем дольше человек учился, тем ниже у него давление? Вероятно, нет. Быть может, образованные люди меньше курят. Может, так, а может, и нет. Быть может, они выкуривают меньше сигарет в день (зависимость от дозировки). Может, так, а может, и нет. Быть может, они начали курить позже, а бросили раньше. Может играть роль питание. Спорт. Условия работы. Генетика. Быть может, существует малая подгруппа, которая работает к крайне нездоровых условиях, что само по себе приводит к повышению давления. Количество потенциальных факторов зашкаливает, причем многие из них по сей день неизвестны и, возможно, будут открыты в будущем: что, например, если включение в диету мандаринов снижает давление? Кроме того, может играть роль комбинация факторов. Быть может, питание существенно, только когда человек не занимается никаким спортом.

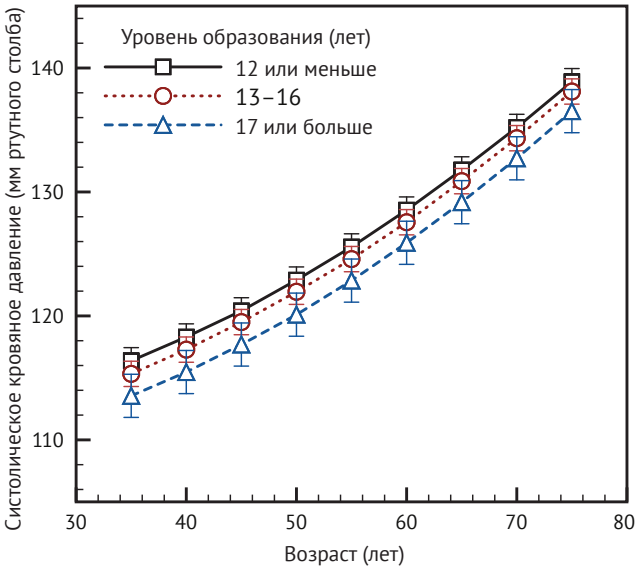


Рис. 3.11. Зависимость среднего систолического давления от возраста для трех генеральных совокупностей людей из городка Фрейминген в штате Массачусетс, США. «Усы» показывают стандартную ошибку каждого среднего. Три совокупности различаются продолжительностью образования. Кровяное давление увеличивается с возрастом. Кроме того, давление ниже всего в группе с наибольшим числом лет образования и выше всего в группе с наименьшим числом лет образования. Что мы можем отсюда заключить? Как показано ниже, не много. Данные заимствованы из работы Loucks et al. [1]

Различие в кровяном давлении между разными с точки зрения образования группами составляет всего 2 мм ртутного столба. Чтобы эта цифра была понятна в контексте, измерьте свое давление и повторите измерение через 5 мин. Вы увидите, что 2 мм ртутного столба – это очень мало по сравнению с вашей персональной изменчивостью (ε_{ij}) и по сравнению с более широким диапазоном групповой изменчивости (v_i). К тому же давление сильно зависит от активности. Быть может, разница существует, только когда давление измеряли в состоянии покоя. А быть может, и нет. Основная проблема таких когортных исследований в том, что имеется слишком много факторов, которые причинно связаны, но не могут контролироваться. Чтобы контролировать все эти эффекты и комбинации, размер выборки должен быть больше, чем проживает людей на планете. Да и вообще, имеет ли смысл исследовать разницу в 2 мм ртутно-

го столба? Если вы хотите снизить свое давление, немного занятий спортом сделают больше и обойдутся куда дешевле тысяч долларов, уплаченных за получение образования.

Следствие 5b. Небольшой размер эффекта. Как показано выше, к исследованиям с небольшим размером эффекта следует подходить с сугубой осторожностью. Однако малый размер эффекта не всегда является проблемой. Во-первых, хорошо бы уменьшить побочные эффекты лекарства, потребляемого миллионами людей, пусть даже всего на 1 %. Во-вторых, многие важные открытия начинались с малых эффектов, и лишь последующие исследования позволили отточить методы и произвести больший эффект.

Следствие 5c. Выводы. Важно, что проблему может вызывать как большой, так и малый размер выборки. Хорошо известно, что выборка *малого* размера составляет проблему из-за недостаточности. В меньшей степени осознано, что и выборка *большого* размера может оказаться проблематичной, если размер эффекта мал, поскольку даже крохотные различия могут стать значимыми. В частности, когортные исследования с малым размером выборки и большим размером эффекта часто бесполезны, поскольку небольшая корреляция между исследуемым фактором и посторонними факторами может привести к значимым результатам. Поэтому важно принимать во внимание как размер эффекта, так и размер выборки. Но если размер выборки n обычно упоминается, то размер эффекта далеко не всегда. Для t-критерия размер эффекта часто выражают в форме d Коэна (см. также главу 4). В следующих главах мы поговорим о размерах эффекта в других критериях.

Как читать статистику? Для разных выборок оценка эффекта d' может сильно разниться. Чем больше размер выборки n, тем меньше дисперсия и тем точнее оценка. Следовательно, первым делом нужно посмотреть, достаточно ли n велико. Если да, решить, отвечает ли размер эффекта предмету исследования. Крохотные размеры эффектов важны лишь в немногих случаях и могут проистекать по причине запутанных, неидентифицируемых факторов. В части III мы увидим, что комбинация размера выборки и размера эффекта может приводить к интересным мыслям по поводу того, стоит ли доверять исследованию. Например, мы зададимся вопросом, насколько вероятно, что четыре эксперимента, в каждом из которых размеры выборки и эффекта были малы, могут привести к значимым результатам с p-значениями ниже 0.05.

Что следует запомнить

1. Даже небольшой размер эффекта может приводить к значимым результатам, если размер выборки достаточно велик.
2. Не сравнивайте *p*-значения двух экспериментов с разными *n*: из того, что *p* меньше, не следует большая значимость.
3. Статистическая и практическая значимость не одно и то же.
4. Отсутствие доказательства еще не есть доказательство отсутствия: избегайте делать выводы из нулевого результата.
5. Не сравнивайте значимый эксперимент с незначимым контрольным экспериментом.
6. Когортные исследования с малым эффектом обычно бесполезны.
7. Утверждение вида «X истинно» действительно истинно, только когда групповая изменчивость равна нулю.

Литература

1. Loucks EB, Abrahamowicz M, Xiao Y, Lynch JW. Associations of education with 30 year life course blood pressure trajectories: Framingham Offspring Study. BMC Public Health. 2011;28(11):139. https://doi.org/10.1186/1471-2458-11-139.

Глава 4

Вариации на тему *t*-критерия

Что вы узнаете из этой главы

В главе 3 мы познакомились с базовой концепцией статистики в контексте ТОС. Здесь мы представим введение в классическую теорию проверки гипотез и опишем вариации на тему *t*-критерия.

4.1. Немного терминологии

Тип эксперимента

- Экспериментальное исследование: образцы случайно распределены между двумя группами. Например, пациенты *случайным образом* включаются либо в экспериментальную группу, получающую потенциально эффективное лекарство, либо в контрольную группу, получающую плацебо.
- Когортное исследование: группы определены заранее заданными метками, например пациенты и контрольная группа, вегетарианцы и мясоеды, космонавты и обитатели земли. Когортные исследования проводятся часто и с пользой, но сталкиваются с рядом серьезных проблем, описанных в главе 3, следствие 5а.

Типы переменных и метрики. На графиках по оси *x* обычно откладывается независимая переменная, а по оси *y* зависимая. Для переменных обоих типов существует четыре основных типа измерительных шкал.

- Номинальная: значения не упорядочены. Например, кровяное давление определяется для людей из разных стран. По оси *x* можно отложить страны в *любом* порядке. Другой пример номинальной шкалы: терапия A и B.
- Порядковая: только ранги. Например, ранг генерала выше, чем лейтенанта, но нельзя сказать, что он вдвое выше. По оси *x* от-

кладываются ранги в порядке *возрастания*. Расстояние между точками на оси *x* не имеет значения.

- Интервальная: значения можно складывать и вычитать, но умножение и деление не имеют смысла. День, когда на термометре 30 °C, жарче, чем день, когда термометр показывает 15 °C, но не вдвое жарче, потому что 0 °C не означает отсутствие тепла. Поэтому физики пользуются шкалой Кельвина, в которой 0 K является абсолютным нулем.
- Относительная: значения можно складывать, вычитать, умножать и делить, в частности, имеют смысл отношения. Классический пример – измерение веса (например, в килограммах). Значение 0 определяет начальную точку шкалы и означает «нет», к чему бы ни относилась переменная: к длине, весу или чему-то еще.

Типы критериев

- Параметрический критерий: критерий, в котором предполагается некоторая модель распределения данных. Например, в главе 3 мы предполагали, что высоты деревьев в генеральной совокупности распределены нормально. Параметрические распределения обычно можно описать небольшим числом параметров (например, средним и стандартным отклонением в случае нормальных распределений).
- Непараметрический критерий: никакое конкретное распределение не предполагается. Некоторые непараметрические эквиваленты *t*-критерия обсуждаются ниже.

4.2. Стандартный подход: проверка нулевой гипотезы

В главе 3 мы объясняли статистику в контексте TOC. Здесь же мы опишем классический подход – проверку нулевой гипотезы – на примере двухвыборочного *t*-критерия.

Шаги принятия статистического решения в двухвыборочном *t*-критерии следующие.

1. Сформулировать альтернативную гипотезу, обозначаемую H_1, например Терапия A лучше Терапии B (или отличается от нее).
2. Предположить, что верна гипотеза H_0: между Терапией A и Терапией B нет различий.
3. На основании имеющихся данных вычислить стандартную ошибку:

$$s_{\bar{X}_A - \bar{X}_B} = s\sqrt{2/n}.$$

4. Вычислить статистику критерия, как в главе 3,

$$t = \frac{\bar{X}_A - \bar{X}_B}{s_{\bar{X}_A - \bar{X}_B}}$$

   и соответствующее p-значение.

5. Принять решение. Если $p \leqslant 0.05$, отвергнуть гипотезу H_0 и принять H: считать эффект значимым. Если $p > 0.05$, нельзя высказать никакого утверждения, в частности нельзя заключить, что H_0 верна.

У описанного подхода есть полезное свойство: устанавливается предельная вероятность допустить ошибку типа I (ложноположительный результат). Предположим, что нулевая гипотеза в действительности верна; это значит, что выборка формируется из распределения чистого шума. Если выбрать много примеров из такого распределения, то обнаружится, что в среднем p меньше 0.05 только в 5 % случаев. Можно усилить ограничение и потребовать, чтобы $p < 0.01$, в таком случае p будет меньше 0.01 только в 1 % случаев. Конечно, не обойтись без компромисса: чем строже ограничение, тем больше частота ошибок типа II (ошибочных пропусков) – когда примеры на самом деле выбираются из распределения альтернативной гипотезы.

4.3. Другие t-критерии

4.3.1. Одновыборочный *t*-критерий

Иногда требуется сравнить одно среднее с фиксированным значением. Это называется одновыборочным t-критерием. Например, исследователь хочет показать, что в результате терапии повышается IQ, в среднем равный 100. Мы предполагаем, что без терапии оценка среднего распределения $\mu_0 = 100$. Следовательно, если нулевая гипотеза верна и терапия не дает никакого эффекта, то мы получим стандартизованное распределение IQ в генеральной совокупности. Стандартное отклонение среднего выборочного распределения равно

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}, \tag{4.1}$$

т. е. стандартной ошибке среднего. *t*-значение вычисляется по формуле

$$t = \frac{\overline{x} - \mu_0}{s_{\overline{x}}}, \tag{4.2}$$

а количество степеней свободы равно

$$df = n - 1. \tag{4.3}$$

Располагая этой информацией, мы можем вычислить *p*-значение и принимать решения так же, как в двухвыборочном *t*-критерии.

4.3.2. *t*-критерий для зависимых выборок

Часто бывает, что имеются две выборки данных, но они как-то связаны между собой. Например, исследователь хочет проверить, повышает ли терапия уровень эритроцитов в крови, для чего сравнивает их концентрацию до и после терапии у одних и тех же обследуемых. Другой пример – мы хотим измерить, какие боевики предпочитают пары, состоящие в отношениях. Ключевая характеристика – каждая отметка, измеренная в одной выборке, может быть однозначно привязана к оценке в другой выборке. Следовательно, для каждой пары мы можем создать *разностный пример*. Итак, два показателя концентрации эритроцитов для данного пациента – до (x) и после (y) терапии – дают один разностный пример для этого пациента:

$$d = y - x. \tag{4.4}$$

Теперь у нас имеется одна выборка разностных примеров, и мы можем применить к ней одновыборочный *t*-критерий, как было описано выше. Стандартная ошибка разностных примеров равна

$$s_{\overline{d}} = \frac{s_d}{\sqrt{n}}, \tag{4.5}$$

где s_d – стандартное отклонение выборочных разностных примеров d (обратите внимание на контекст обсуждения и не путайте эту величину с d Коэна). Как и прежде, мы можем сравнить выборочное среднее с гипотетической разностью средних генеральных совокупностей (только нужно обязательно брать разность средних генеральных совокупностей так же, как для индивидуальных примеров):

$$t = \frac{\overline{d} - (\mu_y - \mu_x)}{s_{\overline{d}}}. \tag{4.6}$$

Вычислим p-значение так же, как раньше, по формуле:

$$df = n - 1. \tag{4.7}$$

Этот критерий также называют t-критерием с повторными измерениями, парным t-критерием или внутрисубъектным t-критерием.

4.3.3. Односторонние и двусторонние критерии

В примерах выше нас неявно интересовало, различаются ли средние двух групп ($\mu_A \neq \mu_B$), т. е. мы не задавались вопросом, лучше ли Терапия А, чем Терапия В. Такие t-критерии называются *двусторонними*, потому что большое t-значение может встретиться в любом хвосте нулевого выборочного распределения. Можно было бы также постулировать, что если различие имеется, то Терапия А лучше Терапии В ($\mu_A > \mu_B$). Наоборот, можно было постулировать, что если различие имеется, то Терапия В лучше Терапии А ($\mu_B > \mu_A$). В этих случаях t-значение может находиться только в одном хвосте (см. рис. 3.5). Таким образом, для одностороннего критерия нужно удовлетворить только одно пороговое условие, а потому для обеспечения желаемой частоты ложноположительных результатов 0.05 требуется меньший порог. Поэтому мощность одностороннего критерия выше, чем двустороннего.

Однако применение одностороннего t-критерия несет в себе противоречие. В рассмотренном в главе 3 примере деревья на северном склоне могут быть выше или ниже деревьев на южном склоне. Следовательно, двусторонний t-критерий подходит лучше, если только нет веских оснований полагать, что северные деревья выше южных. Односторонний критерий нельзя использовать, когда двусторонний критерий привел к незначимому результату (см. раздел 11.3.5)! Кроме того, не следует делать выбор в пользу одностороннего критерия только потому, что в наблюдаемых данных одно среднее больше другого. Решение о том, какой критерий – односторонний или двусторонний – использовать, следует принимать, только если оно теоретически обосновано, оно не должно основываться на данных или результате вычислений.

4.4. Предположения в основе t-критерия и их нарушения

В традиционных учебниках написано, что основной смысл t-критерия – контроль частоты ошибок типа I (частоты ложных тревог). Отклонения от сформулированных ниже предположений почти

всегда изменяют соответствующую частоту ошибок типа I – иногда сильно, а иногда слабо.

4.4.1. Данные должны быть независимы и одинаково распределены

Выборочные данные должны быть независимы и одинаково распределены (англ. *IID*). Это требование обязательно для многих статистических критериев. Например, мы хотим проверить, верно ли, что болеутоляющее не только уменьшает головную боль, но также снижает температуру тела. Мы можем набрать выборку участников, измерить их температуру до и после приема лекарства и вычислить парный *t*-критерий. Важно, что действия с каждым участником производятся только один раз. Если вы хотите выдвинуть гипотезу о генеральной совокупности, то не можете поставить эксперимент только на себе 10 раз подряд или в течение 10 дней, потому что такие данные не будут независимыми. Быть может, вы единственный человек на планете, на которого данное болеутоляющее действует.

Рассмотрим еще один пример. Измеряется острота зрения в восьми точках поля зрения для трех пациентов. Таким образом, имеется 24 измерения, но они не являются независимыми, поэтому мы не можем выполнить для них *t*-критерий. Можно было бы усреднить все восемь точек для одного пациента и вычислить *t*-критерий. А это значит, что размер выборки равен всего 3, а не 24.

Данные должны быть одинаково распределены, т. е. выбирать их следует из одного и того же распределения генеральной совокупности. Например, в одну выборку нельзя включать высоты растений разных видов. Даже если оба распределения нормальные, дисперсии для дубов и эдельвейсов могут сильно различаться. Если, например, мы измеряем высоты в выборке растений, произрастающих на северном и южном склоне, то различия могут быть велики просто потому, что на севере растет больше дубов и меньше эдельвейсов, чем на юге.

4.4.2. Распределения генеральной совокупности нормальные

Для применения *t*-критерия требуется, чтобы распределения генеральной совокупности были нормальными[1] или чтобы размер выборки был велик (часто значения $n = 30$ достаточно). Однако

[1] Выполнение предположения о нормальности можно проверить с помощью критерия Колмогорова–Смирнова.

t-критерий является довольно робастным для генеральных совокупностей, не сильно отличающихся по форме от нормального. Под робастностью мы понимаем, что частота ошибок типа I близка к желательной (например, 5 %, когда гипотеза H_0 отвергается, если $p < 0.05$). При условии одномодальности[1] распределения даже сильная асимметрия слабо влияет на частоту ошибок типа I *t*-критерия (у асимметричного распределения один хвост длиннее другого).

4.4.3. Шкала зависимой переменной

Поскольку *t*-критерий сравнивает средние, требуется, чтобы зависимая переменная соответствовала шкале измерения. Вычисление среднего для номинальных данных не имеет смысла. Вычисление дисперсии (или стандартного отклонения) для номинальных или порядковых данных тоже бессмысленно. Поскольку при вычислении *t*-критерия используются выборочное среднее и выборочное стандартное отклонение, ни номинальные, ни порядковые данные с его помощью анализировать нельзя.

Существуют разные мнения о том, следует ли применять t-критерий к интервальным данным. Строго говоря, свойства t-критерия требуют относительной шкалы измерений, но во многих случаях его поведение достаточно разумно и для интервальных данных.

Таблица 4.1. Частоты ошибок типа I для 10 000 смоделированных *t*-критериев с разными стандартными отклонениями генеральной совокупности и размерами выборок

	$n_1 = n_2 = 5$		$n_1 = 5, n_2 = 25$	
	$\sigma_2 = 1$	$\sigma_2 = 5$	$\sigma_2 = 1$	$\sigma_2 = 5$
$\sigma_1 = 1$	0.050	0.074	0.052	0.000
$\sigma_1 = 5$	0.073	0.051	0.383	0.47

4.4.4. Равные дисперсии генеральной совокупности

Стандартный двухвыборочный *t*-критерий предполагает, что дисперсия обеих генеральных совокупностей одинакова. Неравные стандартные отклонения, особенно в сочетании с неравными размерами выборок, могут очень сильно повлиять на частоту ошибок

[1] Кривая одномодального распределения имеет только одну вершину. Например, таковым является нормальное распределение. Распределения, имеющие две вершины, называются двухмодальными.

типа I. В табл. 4.1 показана частота ошибок типа I для 10 000 смоделированных *t*-критериев, когда нулевая гипотеза в действительности была верна. Для каждого смоделированного критерия программа генерировала «данные» из распределений генеральных совокупностей и вычисляла для них *t*-критерий. В разных операциях моделирования стандартные отклонения генеральных совокупностей были либо равны (например, $\sigma_1 = \sigma_2 = 1$), либо не равны (например, $\sigma_1 = 5$, $\sigma_2 = 1$), и размеры выборки тоже либо равны (например, $n_1 = n_2 = 5$), либо не равны (например, $n_1 = 5$, $n_2 = 25$).

Таблица 4.1 показывает, что если размеры выборок равны, то различие в стандартных отклонениях генеральной совокупности несколько увеличивает частоту ошибок типа I. Приблизительно для 7 % примеров нулевая гипотеза отвергалась. Однако если размеры выборок не равны и дисперсии тоже различны, то частота ошибок типа I либо много меньше, либо много больше. Если малая выборка сочетается с малым стандартным отклонением генеральной совокупности, то частота ошибок типа I гораздо меньше желаемого порога, 0.05. В этом конкретном наборе моделирований ни один *t*-критерий не отверг нулевую гипотезу. С другой стороны, если малая выборка сочетается с большим стандартным отклонением генеральной совокупности, то частота ошибок типа I приближенно равна 40 %, что примерно в восемь раз больше желаемого порога 5 %! Проблема в том, что по умолчанию *t*-критерий объединяет стандартное отклонение каждой выборки, чтобы породить одну оценку стандартного отклонения генеральной совокупности. Если малая выборка сочетается с малым стандартным отклонением генеральной совокупности, то объединенная оценка слишком велика, и критерий вряд ли отвергнет нулевую гипотезу. Если малая выборка сочетается с большим стандартным отклонением генеральной совокупности, то объединенная оценка слишком мала, и вероятность, что критерий отвергнет нулевую гипотезу, слишком велика.

Эти проблемы можно решить, воспользовавшись вариантом *t*-критерия, который называется критерием Уэлча. Однако это обходится не бесплатно: если стандартные отклонения действительно равны, то мощность критерия Уэлча оказывается меньше, чем стандартного *t*-критерия (меньше вероятность отвергнуть нулевую гипотезу, когда имеется реальное различие).

4.4.5. Фиксированный размер выборки

Перед тем как начинать эксперимент, необходимо зафиксировать размер выборки для обеих групп. По ходу эксперимента изменять

размеры выборок нельзя. Удовлетворить это требование труднее, чем может показаться. Как оно может быть нарушено и что из этого следует, мы обсудим в разделе 10.4.[1]

Таблица 4.2. Параметрические критерии и соответствующие им непараметрические критерии

Параметрический	Непараметрический
Одновыборочный *t*-критерий	Критерий знаков
Двухвыборочный *t*-критерий	Критерий суммы рангов Уилкоксона
t-критерий с повторными измерениями	U-критерий Манна–Уитни

4.5. Непараметрический подход

Если распределение данных не нормальное, то можно рассмотреть применение непараметрического критерия. Каждому из описанных выше *t*-критериев соответствует непараметрический критерий (см. табл. 4.2).

Мощность непараметрических критериев меньше, потому что они не могут пользоваться моделью, т. е. непараметрическим критериям для получения значимых результатов обычно нужны выборки большего размера.

Вычисления для непараметрических критериев сильно отличаются от вычисления *t*-критерия, но они следуют тем же базовым принципам TOC.

Что следует запомнить

1. Для применения *t*-критерия данные должны быть независимы и одинаково распределены, а шкала по оси *y* должно быть относительной.
2. Данные должны быть распределены нормально, или *n* должно быть велико.

[1] Можно зафиксировать размер выборки *n* и применить дополнительное условие, например: всего в выборке 20 участников, но если проверка зрения, проведенная до эксперимента, показала, что у участника пониженная острота зрения, то данный участник на этом этапе исключается и может быть заменен другим.

4.6. Принципиальные основы статистических критериев

Вернемся к *t*-критерию. В своем исследовании мы задались вопросом о том, различаются ли *средние* высоты деревьев. Затем мы сделали допущение о статистической модели – предположили, что высоты деревьев распределены нормально. Из этой модели мы вывели уравнение для *t*-значения, которое называется *статистикой критерия*, и это позволило нам вычислить *p*-значение и тем самым контролировать частоту ошибок типа I. Этот принцип можно применить ко многим статистическим вопросам. Например, можно спросить, различаются ли *дисперсии* двух распределений генеральной совокупности, различаются ли формы распределений генеральной совокупности (критерий χ^2) или отлично ли от 1 отношение двух средних (*z*-критерий). Более сложные критерии вычисляют, например, средние, зависящие от других переменных, а еще более сложные предполагают значительно более сложные модели, например иерархического распределения вероятностей.

Для всех критериев принцип один и тот же, и все их можно понять в рамках TOC, выдвигая те же аргументы, что и для *t*-критерия. Единственное различие заключается в том, что используется статистическая модель, отличная от *t*-распределения. Как именно вычисляются статистики различных критериев, не так существенно для понимания, потому что эти вычисления можно поручить компьютеру. Для всех параметрических критериев *p*-значение объединяет размер эффекта и размер выборки.

4.7. Что дальше?

Всегда полезно планировать максимально простой эксперимент, чтобы можно было применить *t*-критерий или соответствующий непараметрический критерий. Однако максимальная простота не всегда достижима. Например, может понадобиться изучить более двух генеральных совокупностей деревьев, и тогда *t*-критерий неприменим. Увеличение числа переменных, например генеральных совокупностей деревьев, приводит к задаче множественной проверки гипотез, которую мы опишем в следующей части книги. Эту задачу можно решить либо статистическими методами, либо благодаря искусному планированию эксперимента (глава 7). Мы рассмотрим наиболее распространенные методы, потому что они включают подход, не оче-

видный из описания *t*-критерия. Хотя существуют и другие критерии, мы не станем объяснять их в этой книге, потому что она посвящена основам статистики и не является полным сводом методов.

В части I мы ввели в рассмотрение многие фундаментальные термины статистики, необходимые пользователям статистических программ. Читатели, которых не интересуют конкретные тесты, описываемые в части II, могут перейти сразу к части III, где эти термины используются для объяснения того, почему в данный момент мы переживаем кризис науки и статистики.

Часть II

Множественная проверка гипотез

Глава 5

Задача множественной проверки гипотез

Что вы узнаете из этой главы

В части I мы рассматривали самое простое статистическое сравнение, а именно сравнение двух средних. Мы описали *t*-критерий, который обладает большой мощностью и, следовательно, является хорошим критерием, рекомендуемым к применению всюду, где возможно. Но иногда сравнения двух средних недостаточно, например если мы хотим сравнить генеральные совокупности деревьев в трех регионах планеты. В таком случае возникает задача множественной проверки гипотез, в которой риск допустить ошибку типа I выше.

В этой главе мы познакомимся с задачей множественной проверки гипотез и представим поправку Бонферрони как один (неоптимальный) из способов ее решения.

5.1. Независимые проверки

Чтобы понять, в чем состоит задача множественной проверки гипотез, рассмотрим следующую ситуацию. Вычисляя один *t*-критерий, мы знаем, что если нулевая гипотеза в действительности верна, то частота ошибок типа I равна $\alpha = 0.05$. По-другому то же самое можно выразить, сказав, что ошибка типа I (ложная тревога) не возникает в 1–0.05 % случаев, если нулевая гипотеза верна. При вычислении двух независимых *t*-критериев вероятность не допустить ложную тревогу равна $0.95^2 = 0.9$. Для 12 сравнений $0.95^{12} = 0.54$. Таким образом, вероятность хотя бы одной ложной тревоги при 12 сравнениях равна $1 - 0.54 = 0.46$. Следовательно, с увеличением числа сравнений ложные тревоги становятся все более и более вероятными (рис. 5.1). В общем случае вероятность хотя бы одной ошибки типа I для m независимых проверок равна:

$$1 - (1 - \alpha)^m, \qquad (5.1)$$

или $1 - (1 - 0.05)^m$ для $\alpha = 0.05$.

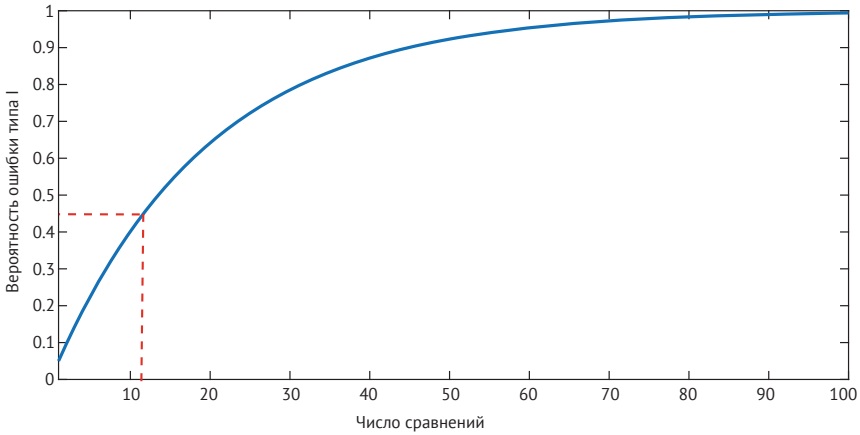


Рис. 5.1. Частота ошибок типа I (частота ложных тревог) сильно зависит от числа сравнений. Так, при 12 сравнениях (красные штриховые линии) вероятность хотя бы одной ложной тревоги равна 0.46 – гораздо выше, чем вероятность 0.05 при одном сравнении

Поправки Бонферрони. Классический способ учесть увеличение частоты ошибок типа I заключается в том, чтобы уменьшить требуемый уровень значимости. Если мы хотим, чтобы частота ошибок типа I для m независимых критериев была равна 0.05, то должны приравнять выражение (5.1) к 0.05 и решить получившееся уравнение относительно α:

$$\alpha = 1 - (0.95)^{\frac{1}{m}} \approx \frac{0.05}{m}. \qquad (5.2)$$

Чтобы частота ложных тревог была равна 0.05 для всех m проверок, необходимо выполнение условия:

$$p < \frac{0.05}{m}. \qquad (5.3)$$

Следовательно, в случае нескольких проверок для достижения требуемой значимости необходимо меньшее p-значение для каждой проверки. Теория обнаружения сигнала говорит, что более консервативный порог всегда связан с компромиссом между правиль-

ными подтверждениями и ложными тревогами. И действительно, при использовании поправки Бонферрони мощность (частота правильных подтверждений) резко уменьшается.

Статистики спорят по поводу того, полезны ли вообще поправки Бонферрони (и аналогичные им), и если да, то в каких случаях. Очевидно, что не следует рассматривать *m* как общее число проверок гипотез, выполненных вами на протяжении всей своей научной карьеры. На самом деле если проверяются гипотезы по совершенно различным вопросам, то кажется разумным заводить отдельную частоту ошибок типа I для каждого вопроса, и поправка вообще не нужна.

Есть еще и такой вариант ситуации, в которой нужна множественная проверка гипотез. Вы набрали выборку под некоторую гипотезу, которая оказалась *не* значимой. Вы решили проверить другие гипотезы. Например, при работе над некоторым заданием не обнаружилось разницы в способности к запоминанию между мужчинами и женщинами. Тогда вы решили проверить, верно ли, что молодые женщины показывают иные результаты, чем пожилые, и то же самое для мужчин. Для каждой из этих гипотез есть риск ложной тревоги, и его нужно скорректировать. Следовательно, задавать слишком много вопросов чревато проблемами. Хотя эти проверки не являются независимыми, поправка Бонферрони может эффективно контролировать частоту ошибок типа I.

5.2. Зависимые проверки

Формула (5.1) справедлива, когда все проверки независимы. Если один и тот же набор данных служит для ответа на много вопросов, то независимость проверок может быть поставлена под сомнение, потому что данные используются многократно. Хотя поправка Бонферрони и способна ограничить частоту ошибок типа I, она может оказаться чрезмерно консервативной, но насколько сильно это повлияет, зависит от природы зависимостей.

Например, допустим, что мы делаем выборку из популяции серебряных карасей в пруду, и нас интересует, правда ли, что больший хвост может служить признаком большего сердца. По чистой случайности в состав выборки попала рыба, по которой можно предположить, что такая зависимость существует, хотя на самом деле в генеральной совокупности ее нет, – наш экземпляр дал ложную тревогу. Теперь мы проверяем вторую гипотезу на той же выборке, а именно что больший хвост является признаком большего объема легких. Предположим, что имеется линейная корреляция между

размером сердца и объемом легких, тогда второй наш анализ тоже даст ложную тревогу.

Пусть мы задаем 10 вопросов о рыбах в пруду. Если нам очень не повезет, то мы будем иметь плохую выборку и 10 неправильных ответов. В общем случае о наличии или отсутствии корреляции между данными мы обычно не знаем, и это еще одна причина воздержаться от предъявления более одного вопроса к выборке.

5.3. Сколько научных результатов неверно?

Как уже отмечалось, частота ошибок типа I обычно задается равной 5 %. Поэтому можно предположить, что 5 % всех научных результатов, полученных с применением классической статистики, неверны. Однако это не так. Такое утверждение было бы истинным, если бы во всех проведенных экспериментах размер эффекта был равен нулю ($\delta = 0$). Однако ученые обычно ищут реально существующие эффекты, поэтому, скорее всего, во многих экспериментах эффект имеет место, и, следовательно, шансов допустить ошибку типа I нет. Предположим, что ученые ставят только такие эксперименты, в которых действительно имеется эффект. В этом случае ошибок типа I нет, поскольку нулевая гипотеза неизменно не верна. Следовательно, количество неверных научных результатов зависит от частоты случаев отсутствия эффекта (см. главу 1). Это число чаще всего неизвестно, и потому мы не знаем, сколько результатов является ложными тревогами (или ошибочными подтверждениями).

Что следует запомнить

1. К набору данных можно предъявлять только один вопрос. В противном случае следует учитывать несколько сравнений.
2. Старайтесь планировать максимально простые эксперименты.
3. Если спланировать простой эксперимент не получается и необходимо выполнять более одного группового сравнения, читайте следующую главу.

Глава 6

Дисперсионный анализ (ANOVA)

Что вы узнаете из этой главы

В главе 3 мы говорили о том, как сравнивать средние двух групп. В этой главе мы изучим, как сравнивать средние более двух групп.

6.1. Однофакторный ANOVA с независимыми переменными

Допустим, мы хотим исследовать влияние географического региона на высоту деревьев. Мы могли бы взять деревья вблизи экватора, на 49-й и на 60-й параллели. Нас интересует, будет ли средняя высота деревьев из всех трех регионов одинакова (рис. 6.1). Поскольку региона три, мы не можем использовать *t*-критерий, потому что он применим только к сравнению двух групп.

В принципе, можно было бы вычислить три *t*-критерия и сравнить все возможные пары средних (экватор с 49-й параллелью, экватор с 60-й параллелью и 49-ю параллель с 60-й). Но в этом случае, как показано в главе 5, мы столкнулись бы с задачей множественной проверки гипотез и присущим ей неприятным эффектом возрастания частоты ошибок типа I при увеличении числа сравнений. Подобные ситуации – повод для применения дисперсионного анализа (analysis of variance – ANOVA), в котором используется остроумный прием, позволяющий уйти от множественной проверки гипотез.

6.2. Логика ANOVA

Терминология

В области однофакторного дисперсионного анализа с *m* группами применяются разные термины:

- путь = фактор,
- группа = комбинация условий = уровень.

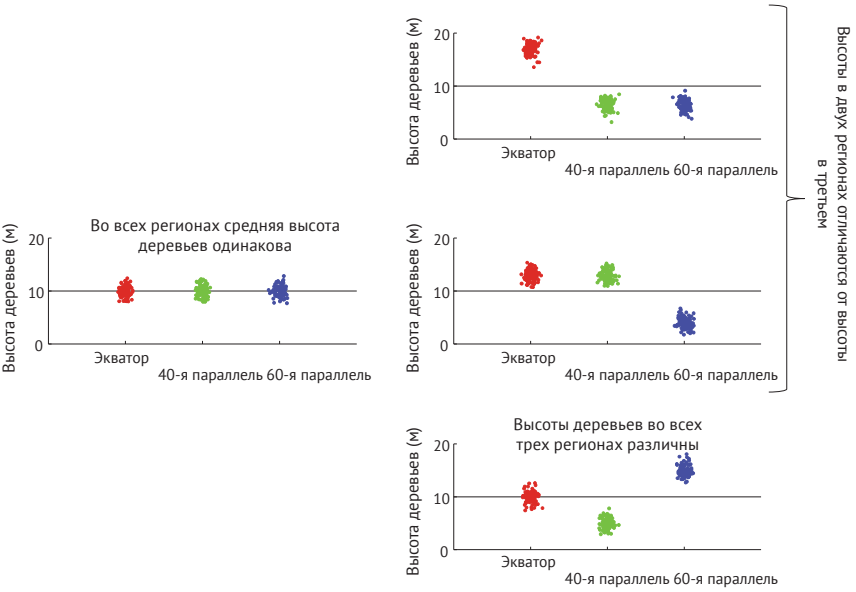


Рис. 6.1. Мы исследуем высоту деревьев на трех разных широтах. Слева: средние высоты на всех трех широтах одинаковы, как показывает горизонтальная прямая. Справа: по крайней мере на одной широте средняя высота не такая, как на двух других. Горизонтальная прямая показывает среднюю высоту деревьев по всем трем группам, она называется «общим средним»

Логика ANOVA проста. Мы упрощаем альтернативную гипотезу, спрашивая, верно ли, что по крайней мере одна из трех генеральных совокупностей деревьев выше других. Следовательно, мы формулируем *одну* гипотезу вместо трех, объединяя альтернативные гипотезы.

Нулевая гипотеза:

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

Объединенные альтернативные гипотезы:

$$H_1 : \mu_1 = \mu_2 \neq \mu_3,$$
$$H_1 : \mu_1 \neq \mu_2 = \mu_3,$$
$$H_1 : \mu_1 \neq \mu_3 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3.$$

В ANOVA, как и в t-критерии, предполагается, что дисперсия генеральной совокупности σ для всех групп одинакова. Если нулевая гипотеза верна, то средние *генеральных совокупностей* объясняются одной лишь дисперсией σ, т. е. случайными различиями высот (шумом), а не систематическими различиями, связанными с географи-

ческим регионом (рис. 6.1). Оказывается, что если нулевая гипотеза верна, то изменчивость средних можно использовать для оценивания σ (путем умножения на размер выборки). В ANOVA эта оценка на основе средних сравнивается с прямой оценкой, вычисленной в каждой группе.

Теперь предположим, что средние высоты деревьев в трех географических регионах на самом деле различны. Тогда высоты индивидуальных деревьев зависят как от внутригрупповой дисперсии σ, так и от изменчивости групповых средних. В таком случае оценка σ, основанная на изменчивости внутри каждой группы, окажется близкой к истинному значению. В ANOVA вычисляется отношение обеих оценочных дисперсий, которое называется F-значением:

$$F = \frac{\text{Оценка дисперсии на основе изменчивости межгрупповых средних}}{\text{Оценка дисперсии на основе изменчивости внутригрупповых средних}}.$$

Формально это выражение можно записать так:

$$F = \frac{\dfrac{\sum_{j=1}^{k} n_j (M_j - M_G)^2}{k-1}}{\sum_{j=1}^{k} \dfrac{\sum_{i=1}^{n_j} (x_{ij} - M_j)^2}{n_j - 1}},$$

где k – число групп (три генеральных совокупности деревьев), n_j – число образцов в группе j (число деревьев внутри каждого географического региона, для которого набиралась выборка), M_j – среднее по группе j (среднее по выборке из географического региона j), M_G – общее среднее по всем вообще образцам, а x_{ij} – i-й пример в группе j (высота одного дерева). Чтобы было проще отличать средние от индивидуальных образцов, мы употребляем символы M_j и M_G вместо традиционного символа для обозначения выборочного среднего $\bar{x}$. Умножение на n_j в числителе служит для назначения отклонениям групповых средних от общего среднего веса, равного числу деревьев в группе, так чтобы число образцов, вносящих вклад в оценку дисперсии, было одинаковым в числителе и в знаменателе.

Рассмотрим два крайних случая. Первый – когда нулевая гипотеза верна (как на рис. 6.2 слева). В этом случае оценки дисперсии в числителе и в знаменателе близки, так что F-значение близко к 1. Далее рассмотрим пример альтернативной гипотезы, когда различия

между высотами деревьев в трех географических регионах велики, а σ очень мало, т. е. высоты деревьев в трех генеральных совокупностях сильно различаются, но внутри одной генеральной совокупности почти равны (рис. 6.2 справа). Зарегистрированная измерениями изменчивость в основном определяется различиями между группами, и F-значение велико.

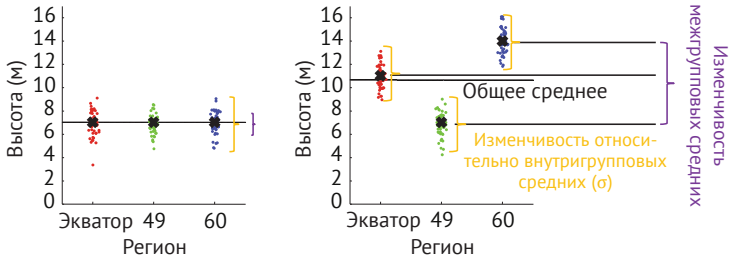


Рис. 6.2. Логика ANOVA. Слева: нулевая гипотеза верна, и, следовательно, средние всех генеральных совокупностей равны. В этом случае вся изменчивость внутригрупповая и может рассматриваться как шум. В ANOVA предполагается, что изменчивость данных одинакова во всех трех генеральных совокупностях деревьев. Справа: нулевая гипотеза неверна. Здесь показан крайний случай, когда изменчивость средних больше, чем изменчивость данных в окрестности среднего. В этом случае большая часть изменчивости данных объясняется влиянием трех разных регионов на высоту деревьев. Обычно ситуация находится где-то посередине между этими двумя крайностями. В ANOVA нулевая гипотеза заключается в том, что все наблюдаемые различия обусловлены шумом. Цель ANOVA – отделить изменчивость, вызванную независимой переменной, от изменчивости данных в окрестности среднего индивидуальной группы

Как и в случае t-критерия, порог статистической значимости выбран так, чтобы частота ошибок типа I была равна желаемой (например, $\alpha = 0.05$). Если F превосходит порог, то мы делаем вывод, что различие значимо (т. е. отвергаем нулевую гипотезу о равенстве межгрупповых средних).

Мы рассмотрели пример однофакторного дисперсионного анализа, в котором имеется один фактор (место произрастания деревьев) и три группы (региона) внутри этого фактора. Группы называются также уровнями. Внутри одного фактора может быть много уровней, т. е. регионов, в которых формируются выборки деревьев. Частным случаем является однофакторный ANOVA с независимыми переменными и двумя уровнями, когда сравниваются два средних, как в t-критерии. На самом деле имеется тесная связь между обоими критериями, и в данном случае $F = t^2$. Здесь p-значение будет одина-

ковым что для ANOVA, что для двустороннего *t*-критерия. Следовательно, ANOVA – обобщение *t*-критерия.

Как и в случае *t*-критерия, число степеней свободы играет важную роль при вычислении *p*-значения. Для однофакторного ANOVA с независимыми переменными и k уровнями имеется два типа степеней свободы, df_1 и df_2. В общем случае $df_1 = k - 1$, $df_2 = n - k$, где n – общее число выбранных образцов во всех группах, например всех деревьев в трех группах. Полное число степеней свободы равно $df_1 + df_2 = n - 1$.

6.3. О чем ANOVA говорит, а о чем нет: апостериорные критерии

Предположим, что, выполняя ANOVA, мы нашли значимый результат. О чем это говорит? Мы отвергаем нулевую гипотезу о том, что все средние равны,

$$H_0 : \mu_1 = \mu_2 = \mu_3,$$

и тем самым принимаем альтернативную гипотезу, что может означать верность любого из следующих утверждений:

$$H_1 : \mu_1 = \mu_2 \neq \mu_3,$$
$$H_1 : \mu_1 \neq \mu_2 = \mu_3,$$
$$H_1 : \mu_1 \neq \mu_3 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2 \neq \mu_3.$$

Отвергая H_0, мы принимаем одну из альтернативных гипотез, но не знаем, какую именно. Это цена за уход от множественной проверки – четыре альтернативных гипотезы сливаются в одну.

Тут ANOVA предлагает второй трюк. Если мы отвергли нулевую гипотезу, то вполне допустимо сравнивать пары средних с помощью так называемых «апостериорных критериев», которые, грубо говоря, соответствуют попарным сравнениям. В отличие от множественной проверки гипотез, обсуждавшейся в главе 5, эти множественные сравнения не приводят к увеличению частоты ошибок типа I, потому что выполняются, только если ANOVA обнаружил главный эффект.

В статистической литературе описано много апостериорных критериев. К числу наиболее распространенных относятся критерии Шеффе, Тьюки и REGW-Q. Этот процесс лучше всего проиллюстрировать на примере, который приводится в конце этой главы.

6.4. Предположения

Предположения ANOVA аналогичны предположениям для *t*-критерия, описанным в главе 4.

1. Независимые выборки.
2. Нормальное распределение генеральной совокупности.
3. Независимая переменная дискретна, а зависимая непрерывна.
4. Однородность дисперсии: дисперсия всех групп одинакова.
5. Размер выборки следует определить до эксперимента и затем не изменять.

6.5. Пример вычисления для однофакторного ANOVA с независимыми переменными

6.5.1. Вычисление ANOVA

Предположим, что проводится турнир по боям на холодном оружии с тремя разными типами оружия: световым мечом, катаной Хатори Ханзо и эльфийским кинжалом (см. рис. 6.3). Вопрос: имеются ли различия в числе побед разным оружием? Нулевая гипотеза заключается в том, что различий нет. Данные и вычисление *F*-значения показаны на рис. 6.3.

Наше окончательное[1] *F*-значение равно 9.14. Это означает, что изменчивость групповых средних относительно общего среднего в 9.14 раз больше изменчивости отдельных данных относительно их группового среднего. Следовательно, большая часть изменчивости проистекает из различия средних, а гораздо меньшая – из изменчивости в пределах каждой генеральной совокупности. *F*-значение 9.14 приводит к *p*-значению 0.0039 < 0.05, и мы заключаем, что результаты значимы, т. е. отвергаем нулевую гипотезу о том, что среднее число побед разным оружием одинаково ($F(2, 12) = 9.14$, $p = 0.0039$). Кроме того, можно заключить, что по крайней мере для одного типа оружия число побед не такое, как для других. Теперь можно использовать один из апостериорных критериев, чтобы выяснить, какой вид (или виды) оружия превосходят остальные.

[1] Если данные анализируются статистической программой, то получится $F = 9.13$. Различие связано с округлением MS_{Within} на рис. 6.3.

Световой меч	$(x_i - M)^2$
6	$(6 - 5)^2 = 1$
8	9
5	0
4	1
2	9
M = 5	SS = 20

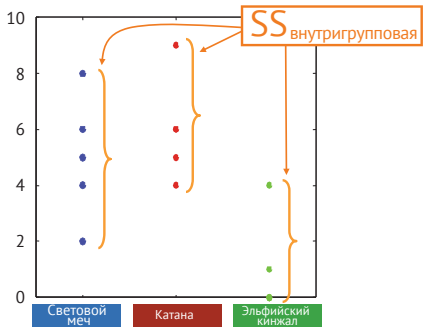
Катана	$(x_i - M)^2$
6	0
5	1
9	9
4	1
6	0
M = 6	SS = 14

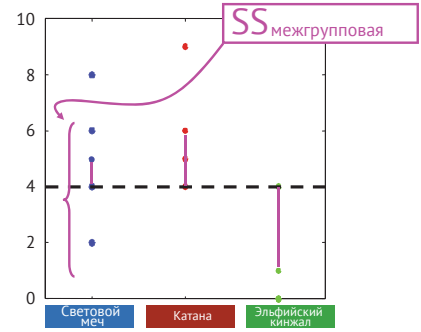
Эльфийский кинжал	$(x_i - M)^2$
0	1
4	9
0	1
1	0
0	1
M = 1	SS = 12



Общее среднее

$$M_G = \frac{\sum_{k=1}^{3}\sum_{i=1}^{n_k} x_{ik}}{N} = \frac{6+8+5+4+2 \;+\; 6+5+9+4+6 \;+\; 0+4+0+1+0}{15} = 4$$





$$SS_{within} = \sum_{k=1}^{3} SS_k$$

$$= 20 + 14 + 12 = 46$$

$$SS_{between} = \sum_{k=1}^{3} n_k (M_k - M_G)^2$$

$$= 5(5 - 4)^2 + 5(6 - 4)^2\,5(1 - 4)^2 = 70$$

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{46}{12} = 3.83$$

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{70}{2} = 35$$

$$F = \frac{MS_{between}}{MS_{within}} = \frac{35}{3.83} = 9.14$$

Источник	SS	df	MS	F	p
Межгрупповые	70	2	35	9.14	0.0039
Внутригрупповые	46	12	3.83		

Рис. 6.3. Пример вычисления для однофакторного ANOVA с независимыми переменными. Каждый из трех видов оружия используется пятью независимыми бойцами, т. е. всего имеется 15 бойцов. Поэтому мы имеем ANOVA типа 1 × 3. Противники в схватках не принадлежат к числу 15 выбранных бойцов. В верхней части рисунка показано, сколько поединков было выиграно различными видами оружия. В нижней таблице показаны вычисленные данные. Например, световым мечом в среднем было выиграно пять поединков. Далее мы вычисляем изменчивость для каждого меча, которая называется также внутригрупповой суммой квадратов (SS). Для этого мы вычитаем каждое значение из среднего и возводим

6.5.2. Апостериорные критерии

Для вычисления апостериорных критериев есть разные процедуры, но для иллюстрации общих принципов мы здесь остановимся на критерии Шеффе.

Идея критерия Шеффе – выполнить несколько сравнений путем вычисления попарных ANOVA (например, световые мечи против катан, световые мечи против эльфийских кинжалов и катаны против эльфийских кинжалов). Одно из предположений ANOVA – что дисперсии всех генеральных совокупностей одинаковы. Если это правда, то наилучшей оценкой изменчивости внутри каждой генеральной совокупности будет объединенная оценка общего ANOVA, вычисленная в виде MS_{within} (в данном случае 3.83). В критерии Шеффе используется также величина $df_{between}$ из общего ANOVA, а вычисления для этого критерия показаны на рис. 6.4.

p-значение для каждого сравнения вычисляется с использованием числа степеней свободы из оригинального ANOVA (т. е. $df_{between} = 2$ и $df_{within} = 12$). В итоге для наших апостериорных критериев получаются результаты, показанные в табл. 6.1. Только второе и третье сравнение оказываются ниже критического порога $\alpha = 0.05$, и, таким образом, можно заключить, что световые мечи отличаются от эльфийских кинжалов ($F(2, 12) = 5.22$, $p = 0.023$) и что катаны отличаются от эльфийских кинжалов ($F(2, 12) = 8.15$, $p = 0.006$), но между световыми мечами и катанами не удалось найти значимого различия ($F(2, 12) = 0.33$, $p = 0.728$).

Рис. 6.3 (*Окончание*) разность в квадрат. Чтобы вычислить внутригрупповую дисперсию, мы складываем все три суммы квадратов. В данном случае получается 46. Следующий важный шаг – вычислить изменчивость средних. Для этого мы сначала находим общее среднее M_G, равное среднему числу побед во всех 15 поединках. В данном примере общее среднее равно 4. Далее для каждого вида оружия среднее вычитается из общего среднего, разности возводятся в квадрат и умножаются на число поединков с данным видом оружия (в этом примере 5). Получается 70. Затем мы делим каждую из двух сумм квадратов на число степеней свободы df_1 и df_2, чтобы получить дисперсии. У нас имеется три вида оружия, поэтому $df_1 = 3 - 1$, так что 70 делится на 2. Было 15 поединков, поэтому $df_2 = 12$, так что 46 делится на 12 (MS означает «mean square» – среднеквадратичное). Для нахождения статистики критерия мы вычисляем частное от деления 35 на 3.83 и получаем $F = 9.14$. Как и в случае *t*-значения, *F*-значение легко вычислить вручную. Что до p-значения, то мы воспользовались статистической программой, которая дает $p = 0.0039$. На выходе статистического пакета программ итоговые результаты вычислений представлены в виде подобных рисунков. В научных публикациях результаты представляются как ($F(2, 12) = 9.14$, $p = 0.0039$)

Световой меч
6
8
5
4
2
Sum = 25
M = 5

Катана
6
5
9
4
6
Sum = 30
M = 6

Эльфийский кинжал
0
4
0
1
0
Sum = 5
M = 1

$df_{between} = 2$

$MS_{within} = 3.83$

Из рассмотренной выше таблицы ANOVA

$$SS_{between} = \sum_{k \in comp} n_k (M_i - G_{comp})^2$$

$$MS_{between} = \frac{SS_{between}}{df_{between}}$$

$$F = \frac{MS_{between}}{MS_{within}}$$

Световой меч против **Катана**

$$G_{comp} = \frac{25 + 30}{10} = 5.5$$

$$SS_{between} = 5(5 - 5.5)^2 + 5(6 - 5.5)^2 = 2.5$$

$$MS_{between} = \frac{2.5}{2} = 1.25$$

$$F = \frac{1.25}{3.83} = 0.3264$$

Световой меч против **Эльфийский кинжал**

$$G_{comp} = \frac{25 + 5}{10} = 3$$

$$SS_{between} = 5(5 - 3)^2 + 5(1 - 3)^2 = 40.0$$

$$MS_{between} = \frac{40}{2} = 20$$

$$F = \frac{20}{3.83} = 5.2219$$

Катана против **Эльфийский кинжал**

$$G_{comp} = \frac{30 + 5}{10} = 3.5$$

$$SS_{between} = 5(6 - 3.5)^2 + 5(1 - 3.5)^2 = 62.5$$

$$MS_{between} = \frac{62.5}{2} = 31.25$$

$$F = \frac{31.25}{3.83} = 8.1593$$

Рис. 6.4. Вычисление сравнений в апостериорном критерии Шеффе для примера с боями на холодном оружии. Сначала вычисляется общее среднее для каждого сравнения (G_{Comp}), равное средней сумме всех примеров из рассматриваемой пар групп. Затем мы вычисляем сумму квадратов отклонений этого общего среднего от групповых средних по двум группам ($SS_{between}$). Далее мы находим величины $MS_{between}$, поделив $SS_{between}$ на $df_{between}$ из описанного выше алгоритма ANOVA (в данном случае 2). Наконец, вычисляется F-значение для каждого сравнения путем деления на величину MS_{within} из описанного выше ANOVA (в данном случае 3.83)

Обычно такие различия иллюстрируются с помощью графика, показывающего среднее число побед для каждого из трех типов оружия с «усами», обозначающими стандартную ошибку каждого среднего. При этом прямые, над которыми располагаются звездочки, соединяют значимо различные виды оружия (рис. 6.5).

Таблица 6.1. Результаты апостериорного критерия Шеффе для трех сравнений

Сравнение	Результат
1 против 2[a]	$F(2,12) = 0.33, p = 0.728$
1 против 3[b]	$F(2,12) = 5.22, p = 0.023$
2 против 3[c]	$F(2,12) = 8.16, p = 0.006$

[a] Световые мечи против катан
[b] Световые мечи против эльфийских кинжалов
[c] Катаны против эльфийских кинжалов

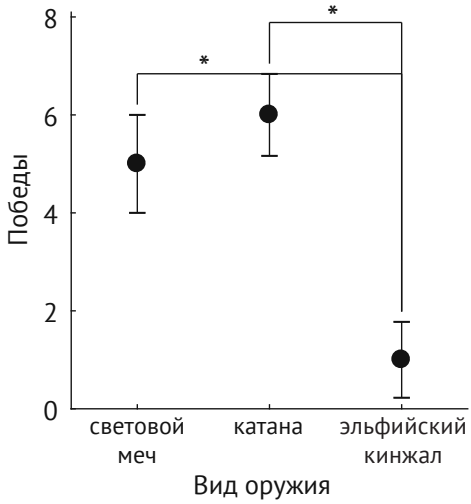


Рис. 6.5. Среднее число выигрышей для всех трех типов оружия со стандартной ошибкой. Линии со звездочкой соединяют значимо различные виды оружия

6.6. Размер эффекта

Как и в случае *t*-критерия, *p*-значение в ANOVA смешивает воедино размер эффекта и размер выборки. Всегда важно учитывать размер эффекта, который в ANOVA обозначается η^2. Он сообщает, какая доля

полной изменчивости зависимой переменной объясняется изменчивостью независимой переменной. Вычисляется он по формуле

$$\eta^2 = \frac{SS_{between}}{SS_{total}},$$

где

$$SS_{between} = \sum_{j=1}^{k} n_j (\bar{x}_j - M_G)^2, \tag{6.1}$$

$$SS_{total} = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - M_G)^2, \tag{6.2}$$

где M_G – общее среднее (т. е. среднее во всем данным). Это отношение говорит, какая доля полной изменчивости данных объясняется изменчивостью групповых средних. Для рассмотренного выше примера размер эффекта $\eta^2 = 0.60$; согласно рекомендациям Коэна (табл. 6.2) такой эффект считается большим.

Таблица 6.2. Рекомендации Коэна по оцениванию размера эффекта

	Малый	Средний	Большой
Размер эффекта	0.01	0.09	0.25

6.7. Двухфакторный ANOVA с независимыми переменными

Однофакторный ANOVA с независимыми переменными хорошо обобщается на случай нескольких факторов. В этом разделе мы обсудим простейший случай – двухфакторный анализ.

Предположим, что вы со своими друзьями обрели в ходе научного эксперимента сверхспособности и готовитесь ступить на тропу войны с преступностью. Вы и ваши друзья-супергерои не хотят, чтобы враги причинили вред дорогим вам людям, поэтому вам нужны маскировочные костюмы. Кроме того, иногда борьба с преступниками будет происходить днем, а иногда ночью. Вы хотите узнать, какой материал (эластан, хлопок или кожа) лучше всего подходит для борьбы с преступностью, притом что эффективность костюма измеряется количеством поборников зла, пойманных героем, носящим костюм, изготовленный из каждого материала. Кроме того, нужно

знать, влияет ли на выбор лучшего материала время суток. Вы раздаете всем друзьям материал для костюма, назначаете время суток и подсчитываете количество пойманных ими преступников. В каждой группе количество друзей различно. В данном случае можно выдвинуть три отдельные гипотезы.

1. H_0: время суток не влияет на количество пойманных преступников.
   H_1: количество преступников, пойманных днем, отличается от количества преступников, пойманных ночью.
2. H_0: материал костюма не влияет на количество пойманных преступников.
   H_1: по крайней мере для одного материала костюма количество пойманных преступников отличается от количества преступников, пойманных в костюмах из других материалов.
3. H_0: влияние времени суток на количество пойманных преступников не зависит от материала костюма.
   H_1: влияние времени суток на количество пойманных преступников зависит от материала костюма.

Первые две нулевые гипотезы относятся к так называемым *основным эффектам*. Обе основные гипотезы в точности такие же, как при вычислении однофакторных ANOVA. Третья гипотеза относится к *взаимодействию* двух факторов, костюма и времени суток; это новый тип гипотезы. Для измерения основного эффекта материала костюма мы берем среднее число преступников, пойманных группой, облаченной в эластан, усредняем по времени суток (отдельно в дневные и ночные часы) и сравниваем с такими же средними для костюмов из хлопка и кожи. Для измерения основного эффекта времени суток мы берем среднее число преступников, пойманных днем, усредняем по костюмам из эластана, хлопка и кожи и сравниваем с такими же средними для преступников, пойманных ночью.

Для анализа взаимодействия мы рассматриваем все группы по отдельности; нас интересует количество преступников, пойманных группами в костюмах из разных материалов, в виде функции от времени суток (днем или ночью). Если имеет место значимое взаимодействие, то влияние времени суток на число пойманных преступников будет зависеть от материала костюма. Обратно: влияние материала костюма на число пойманных преступников будет зависеть от того, в какое время суток герои борются с преступниками.

Для проверки этих трех нулевых гипотез нужны три отдельные *F*-статистики. В каждой *F*-статистике знаменатель будет таким же,

как в однофакторном ANOVA (т. е. объединенная дисперсия данных о групповых средних, MS_{within} в обозначениях рис. 6.3), но числители ($MS_{between}$) будут зависеть от конкретной проверяемой гипотезы.

На рис. 6.6 приведен пример исходных данных и средних, сравниваемых для всех трех гипотез (см. боковые столбцы слева и справа). Объединение по времени суток показывает, что материал костюма очень слабо влияет на эффективность борьбы с преступностью. Объединение по материалу костюма показывает, что и время суток влияет на эффективность слабо. И лишь при рассмотрении каждого среднего по отдельности мы видим истинное влияние времени суток и материала костюма на количество пойманных нашими друзьями преступников (рис. 6.7). Взаимодействие таково, что эластан лучше днем, а кожа ночью, тогда как хлопок находится где-то посередине.

	Эластан	Хлопок	Кожа	Средние по времени суток
День	18	10	3	9.7
	10	8	5	
	16	12	1	
	12	6	7	
	14	14	9	
Среднее по дневному времени	14	10	5	
Ночь	5	6	15	10
	7	14	13	
	3	10	17	
	9	8	11	
	1	12	19	
Среднее по ночному времени	5	10	15	
Среднее по костюмам времени	9.5	10	10	Общее среднее = 9.8

Рис. 6.6. Количество преступников, пойманных каждым супергероем, а также средние основных эффектов (время суток и материал костюма) и средние по отдельным комбинациям (эластан днем, эластан ночью, хлопок днем и т. д.). Общее среднее вычисляется по всем данным

Этот пример иллюстрирует ценность двухфакторного анализа. Если бы мы выполняли только однофакторный ANOVA, изучая связи между материалом костюма и числом пойманных преступников или между временем суток и числом пойманных преступников, то нашли бы только очень слабые эффекты или вообще никаких. Вклю-

чение же обеих переменных вскрывает истинную природу эффектов и показывает, что эффект одного фактора зависит от уровня другого. На рис. 6.8 показано три возможных результирующих паттерна, которые выделяют лишь один значимый эффект (основной эффект A, основной эффект B, взаимодействие), не учитывая остальные. Можно также изучать комбинации основных эффектов и эффектов взаимодействия.

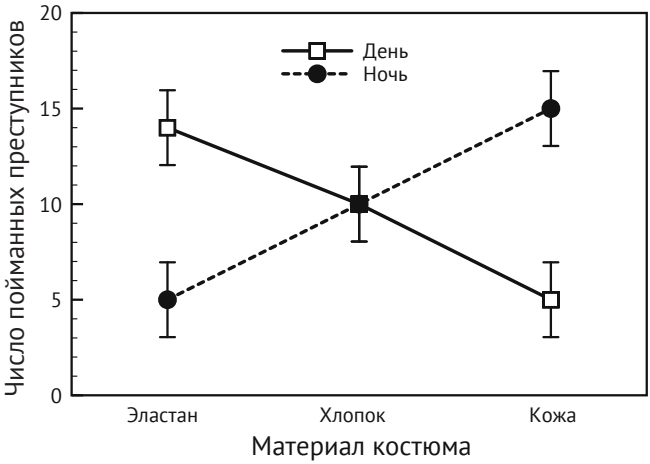


Рис. 6.7. Диаграмма взаимодействия для зависимости числа пойманных преступников от материала костюма и времени суток. Видно, что влияние материала костюма на число пойманных преступников сильно зависит от времени суток

Еще одно достоинство двухфакторного анализа по сравнению с однофакторным – то, что изменчивость, которая в противном случае была бы включена в член ошибки (т. е. MS_{within}), теперь частично объясняется другим фактором, уменьшая тем самым MS_{within} и увеличивая способность обнаруживать присутствующие эффекты.

Таким образом, может показаться, что чем больше мы добавим факторов, тем лучше будем понимать данные и получать значимые результаты. Однако это неверно, потому что теряется мощность каждого добавленного фактора, так как мы имеем меньше примеров, дающих вклад в каждое среднее. Как правило, при увеличении числа факторов нужно увеличивать и размер выборок.

Важно, что если нам удалось выявить значимое взаимодействие, то основной эффект меняется в зависимости от другого фактора. Следовательно, не стоит делать выводов относительно основного эффекта, если имеет место взаимодействие.

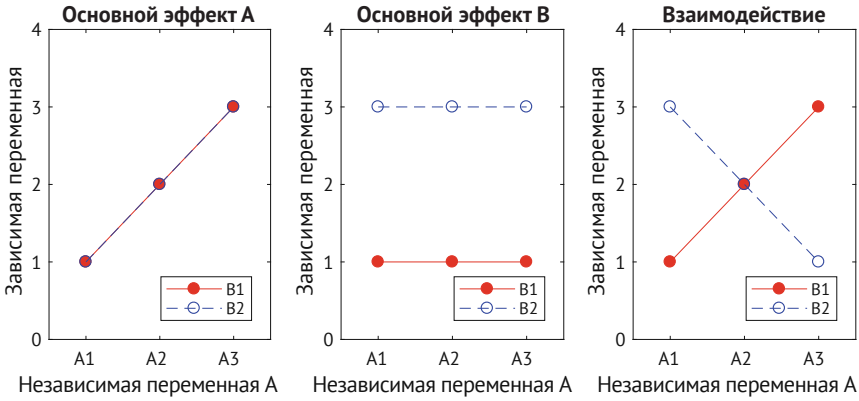


Рис. 6.8. Результаты двухфакторного ANOVA могут выявить три общих типа значимых результатов: основной эффект переменной A, основной эффект переменной B и взаимодействие между A и B. В следующем примере показан пример ANOVA типа 2×3, где A1, A2, A3 могут обозначать материал костюма супергероя (эластан, хлопок, кожа), а B1 и B2 – время суток (ночь, день). Зависимой переменной является число пойманных преступников. Слева: основной эффект A. Материал костюма имеет значение. Больше преступников было поймано героями, облаченными в кожу, а не хлопок. Время суток роли не играет. Днем было поймано столько же преступников, сколько и ночью. Поэтому B1 и B2 расположены друг над другом. В центре: основной эффект B. Материал костюма не имеет значения, а время суток имеет. Днем поймано больше преступников. Справа: взаимодействие, как на рис. 6.7. При наличии значимого взаимодействия значимые основные эффекты взаимодействия обычно не исследуются, потому что анализ влияния одной переменной при различных уровнях другой является более релевантным сравнением

Однофакторный ANOVA позволяет избежать проблему множественной проверки гипотез. Но при многофакторном она возникает снова, немного в ином виде. Например, рассмотрим многофакторный ANOVA типа 2×2 с уровнем значимости 0.05. Для истинно нулевого набора данных (когда средние всех четырех генеральных совокупностей равны) вероятность получить хотя бы одно $p < 0.05$ среди двух основных эффектов и взаимодействия составляет 14 %. Если вы применяете ANOVA для *исследования* набора данных с целью выявления значимых результатов, то должны понимать, что у такого подхода частота ошибок типа I будет выше, чем вы хотели.

Типичная статистическая программа выводит результаты двухфакторного ANOVA в виде, показанном в табл. 6.3.

6.8. ANOVA с повторными измерениями

Рассмотренные выше варианты ANOVA являются прямолинейным обобщением *t*-критерия с независимыми выборками. Существует также обобщение *t*-критерия с зависимыми выборками, которое называется ANOVA *с повторными измерениями*. Такой тип дисперсионного анализа используется, например, когда некоторый аспект состояния здоровья пациента измеряется до, во время и после лечения. В таком случае для одних и тех же пациентов производится три измерения. Мощность ANOVA с повторными измерениями выше, чем у ANOVA с независимыми измерениями, потому что показатели одного пациента сравниваются до сравнения показателей разных пациентов, что уменьшает изменчивость данных. Пример вывода результатов ANOVA с повторными измерениями приведен в табл. 6.4.

Таблица 6.3. Вывод типичной статистической программы для двухфакторного ANOVA

Источник	SS	df	MS	*F*	*p*	η^2
Материал костюма	1.67	2	0.83	0.083	0.920	0.0069
Время суток	0.83	1	0.83	0.083	0.775	0.0035
Костюм × время	451.67	2	225.83	22.58	0.000003	0.6530
Ошибка	240.00	24	10.00			

Столбцы: источник изменчивости, суммы квадратов (*SS*), число степеней свободы (*df*), среднеквадратичное (MS), *F*-значения (*F*), *p*-значения (*p* иногда обозначается «Sig.») и размер эффекта (η^2). В строке «*Ошибка*» представлены результаты вычислений внутрисубъектной изменчивости, а в остальных строках – межсубъектная изменчивость для основных эффектов и взаимодействий

Таблица 6.4. Вывод типичной статистической программы для ANOVA с повторными измерениями

Источник	*SS*	*df*	MS	*F*	*p*	η^2
Межвременная	70	2	35	70	0.00000009	0.94
Внутривременная	40	12				
Межсубъектная	36	4				
Ошибка	110	14				

Здесь представлен примеры симптомов пациента, измеренных до, во время и после лечения (т. е. в разные моменты времени). В первой строке (*«Межвременная»*) показано влияние времени измерения на симптомы. В строке *«Внутривременная»* показана изменчивость субъектов, измеренных в один и тот же момент времени. Она разбита на устойчивые тренды для каждого отдельного субъекта (*«Межсубъектная»*) и случайную ошибку, вызванную такими вещами, как скорость диффузии лекарственного вещества (*«Ошибка»*). Столбцы: источник изменчивости, сумма квадратов (*SS*), число степеней свободы (*df*), среднеквадратичное (MS), *F*-значения (*F*), *p*-значения (*p* иногда обозначается «Sig.») и размер эффекта (η^2). В строке *«Ошибка»* показаны результаты вычислений члена ошибки, который используется в знаменателе *F*-значения. В случае ANOVA с независимыми измерениями этот член совпадает с внутригрупповым. Здесь же мы исключаем из внутригруппового члена изменчивость, обусловленную субъектами, поэтому мы называем остаточную изменчивость просто «изменчивостью ошибки». В остальных строках показана межсубъектная изменчивость для основных эффектов и взаимодействий. Подводя итог этим результатам, мы сказали бы, что имеется значимое влияние момента измерения на симптомы $F(2, 14) = 70$, $p = 0.00000009$. Здесь мы взяли число степеней свободы из строк *«Межвременная»* и *«Ошибка»*, а также *F*- и *p*-значения из строки *«Межвременная»*

Что следует запомнить

1. Дисперсионный анализ (ANOVA) помогает уйти от задачи множественной проверки гипотез – до некоторой степени.
2. Увеличение числа факторов может как повысить, так и понизить мощность.

Глава 7

Планирование эксперимента: подгонка модели, мощность и сложные планы

Что вы узнаете из этой главы

Дисперсионный анализ – один из способов справиться с проблемой множественной проверки гипотез. Гораздо более простой способ – вообще избежать ее путем продуманного планирования эксперимента. Даже если необходимо измерять много переменных, нет нужды подавать их все на вход статистического критерия. В этой главе мы покажем, что, объединив много данных в одну осмысленную переменную или просто опустив какие-то данные, мы сможем увеличить статистическую мощность. Результаты простых и упрощенных экспериментов проще интерпретировать, тогда как для сложных экспериментов с этим могут возникнуть проблемы. Мы также покажем, как вычислить мощность эксперимента, что важно, например, для определения размера выборки.

7.1. Подгонка модели

При сравнении двух средних *t*-критерий обладает высокой мощностью и легко поддается интерпретации. Эксперименты с большим числом групповых сравнений подвержены проблеме множественной проверке гипотез. Чем больше сравнений мы производим, т. е. чем больше имеется групп или уровней, тем ниже мощность. Кроме того, результаты экспериментов с большим числом групп труднее анализировать из-за возможности взаимодействий, которых нет в простых *t*-критериях (глава 6). Поэтому обычно следует предпочесть простой план эксперимента. Однако иногда сложность неизбежна. Классический пример – эксперимент по обучению, когда ка-

чество нужно измерять в большом числе временных точек. Скажем, обучение участников на простой визуальной задаче может быть организовано в виде 10 блоков по 80 испытаний в каждом. Для каждого блока мы определяем процент верных ответов (рис. 7.1a) и смотрим, как повышается качество при переходе от одного блока к другому. Как можно количественно охарактеризовать успешность обучения и вычислить статистики? Нулевая гипотеза заключается в том, что никакого обучения не произошло, т. е. качество во всех 10 блоках одинаково. Интуитивно кажется, что можно было бы использовать ANOVA с повторными измерениями и 10 уровнями, по одному для каждого блока. Однако эта идея неудачна, потому что, во-первых, ANOVA имеет дело с номинальными переменными, т. е. порядок блоков не играет роли. Во-вторых, если бы качество повышалось на протяжении первых пяти блоков, а затем снижалось, то ANOVA показал бы значимый результат, когда качество в блоке 5 значимо отличается от качества в блоке 1. Но такой результат свидетельствует не об обучении, а о странной комбинации приобретения и последующей утраты навыков. В-третьих, мы заметно проиграли бы в мощности. И что же делать? Ниже мы покажем, что вовсе не обязательно подвергать все данные статистическому анализу.

Как показано на рис. 7.1b для эксперимента по обучению, один из возможных подходов – отбросить все блоки, кроме первого и последнего (промежуточные блоки существенны для эксперимента, потому что способствуют обучению, но для статистического анализа не важны). Нулевая гипотеза заключается в том, что качество в этих двух блоках не отличается. Для проверки этой гипотезы мы можем использовать *t*-критерий с повторными измерениями. Однако данные для обучения часто зашумлены, и потому такая процедура приводит к потере мощности. Чтобы уменьшить зашумленность данных, мы можем усреднить первый и последний блок и подать оба средних на вход *t*-критерия с повторными измерениями (рис. 7.1c).

В обоих случаях мы отбрасываем большое количество данных и потому не используем имеющиеся данные в полной мере. Можно было бы поступить лучше, подогнав модель к данным (иначе говоря, аппроксимировать данные моделью). Например, из предыдущих экспериментов мы можем знать, что в результате обучения качество повышается линейно, и смоделировать это формулой $mx + b$, где m – угловой коэффициент кривой обучения, b – координата точки пересечения с осью y (свободный член), а x – номер блока. Для вычисления оптимальных значений параметров m и b отдельно для каждого обучаемого можно воспользоваться компьютерной программой.

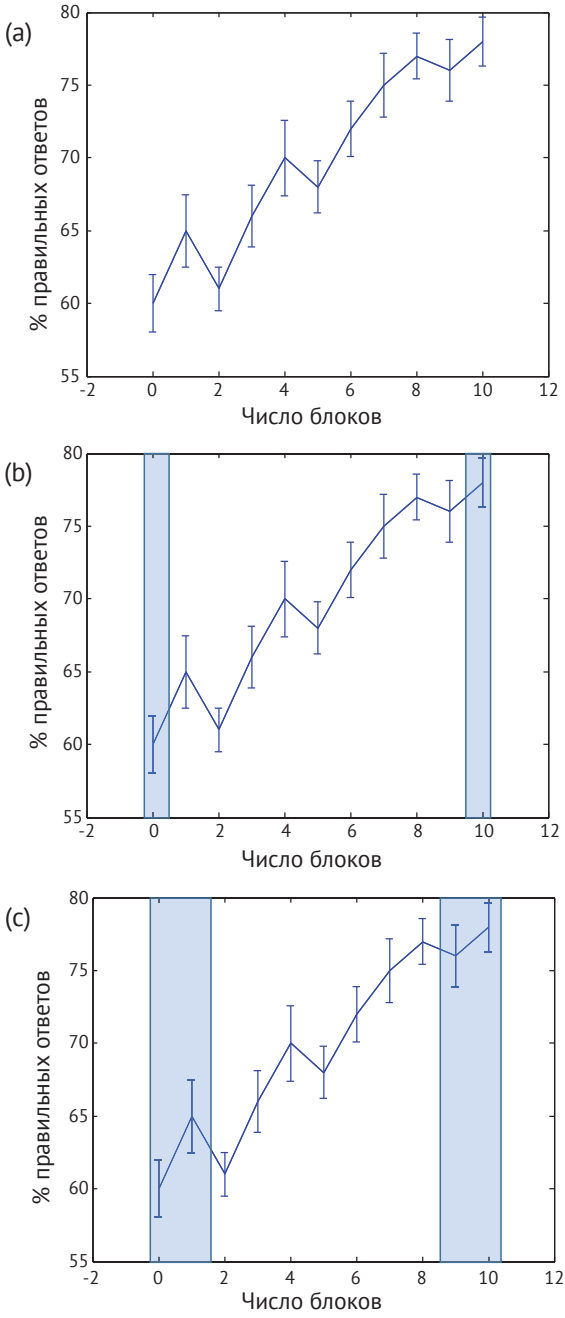


Рис. 7.1. Анализ данных об обучении. (a) Качество улучшается с ростом числа блоков. (b) При статистическом анализе сравниваются только первый и последний блоки. (c) Альтернативно можно усреднить два первых и два последних блока, а затем сравнить средние

Поскольку нас интересует только угловой коэффициент, b можно отбросить. Нулевая гипотеза формулируется так: $m = 0$. Таким образом, для каждого обучаемого мы получаем одно значение m. Если в эксперименте участвовало 12 обучаемых, то мы вычисляем одновыборочный t-критерий по этим 12 значениям m и смотрим, верно ли, что они значимо отличны от 0.

Этот подход обладает большой гибкостью. Например, если обучение описывается не линейной, а экспоненциальной функцией, то можно аппроксимировать данные экспонентой, у которой тоже есть параметр «наклона». Если нас интересует циклический процесс, например изменение температуры в течение суток или количество насекомых на протяжении года, то можно аппроксимировать данные синусоидой. В общем случае можно аппроксимировать данные любой функцией и определить один или несколько ее параметров. Таким образом, мы используем все данные и не теряем мощность. Как действовать – решать экспериментатору. Но свой выбор он должен сделать до начала эксперимента. Нельзя сначала изучить данные, а затем пробовать разные варианты, пока не будет получен значимый результат (см. раздел 11.3.5).

В примере выше показано, как можно упростить статистику, уменьшив число переменных. Мы видим, что необязательно подвергать статистическому анализу все имеющиеся данные в исходной форме. Не существует общих правил упрощения данных, потому что все эксперименты различны. Однако всегда стоит задуматься над тем, на какой главный вопрос должен ответить эксперимент. А затем решить, какие переменные лучше всего подходят для получения ответа и как вычислить статистику. Чем проще план эксперимента и чем меньше переменных, тем лучше.

7.2. Мощность и размер выборки

7.2.1. Оптимизация плана

Часто для проведения эксперимента требуется много усилий и ресурсов. Поэтому обычно имеет смысл заранее оценить, имеет ли эксперимент шансы на успех, и определить, при каких размерах выборки вероятность успеха высока (если целью является обнаружение какого-то эффекта). В общем случае успех означает получение большого t-значения и значимого результата. Сделать это можно двумя способами.

Во-первых, попробуйте увеличить размер эффекта на генеральной совокупности $\delta = (\mu_1 - \mu_2)/\sigma$. Это можно сделать, рассмотрев

ситуации, в которых различие между средними генеральной совокупности ожидаемо велико. Например, можно сначала попытаться найти оптимальные стимулы для визуального эксперимента или тесты в клинических испытаниях, обладающие наибольшей различительной способностью.

Кроме того, попытайтесь уменьшить σ. Значения δ, а значит, t и p, определяются отношением разности средних генеральной совокупности к стандартному отклонению. Можно попробовать уменьшить шумы в измерительных устройствах, ежедневно калибруя их. Можно попробовать сделать выборку более однородной: тестировать пациентов каждый день в одно и то же время, по возможности уравнять количество выпиваемого ими кофе, поручать проведение тестов одному и тому же экспериментатору и т. д. Можно подумать об исключении некоторых пациентов, например, введя ограничения по возрасту, чтобы не путать нарушения в результате болезни с возрастными эффектами. Однако такая стратификация уменьшает общность вашего исследования (см. главу 3, следствия 4). Существует много способов уменьшить σ, и подумать на эту тему всегда стоит.

Во-вторых, увеличьте размер выборки, n. Даже если δ оказывается малым, достаточно большая выборка может дать большое t-значение. При достаточно большой выборке даже малые различия между средними (зашумленный сигнал) можно отличить от ситуации, когда между средними вообще нет различий (чистый шум). Отметим, что этот подход оправдан, только если вы уверены, что даже малый размер эффекта заслуживает внимания. Не имеет смысла набирать и обрабатывать большую выборку только для того, чтобы обнаружить пренебрежимо малый эффект (см. главу 3, следствия 1 и 2).

7.2.2. Вычисление мощности

Даже когда $\delta \neq 0$, эксперимент не всегда дает значимые результаты из-за недостаточности выборки (глава 3). Сейчас мы покажем, насколько вероятно, что для данного $\delta \neq 0$ и данного размера выборки n получается значимый результат. И наоборот, мы покажем, насколько велико должно быть n, чтобы получить значимый результат с заданной вероятностью.

Вероятность успеха эксперимента мы будем оценивать путем вычисления мощности. Мощность – это частота правильных подтверждений. Иначе говоря, это вероятность того, что случайная выборка позволит правильно отвергнуть нулевую гипотезу. Предполагается, что нулевая гипотеза неверна, т. е. имеет место ненулевой эффект. Как показано в главе 3, для вычисления мощности

требуется определенный стандартизированный размер эффекта в генеральной совокупности. Откуда берется этот размер эффекта, зависит от ситуации. Иногда его можно оценить на основании других исследований, в которых ранее изучалось то же (или похожее) явление. А иногда вывести из математических моделей, которые предсказывают поведение в новой ситуации. Вместо того чтобы предсказывать размер эффекта, иногда имеет смысл идентифицировать значение, которое непременно будет представлять интерес или иметь практическую ценность.

После того как размер эффекта на генеральной совокупности определен, мы обращаемся к компьютерным программам, которые вычисляют мощность (никаких простых формул не существует). На рис. 7.2 показан вывод бесплатной программы G*Power. Здесь мы выбрали *t*-критерий из списка *Test family* (Семейство критериев), а в качестве статистического критерия (*Statistical test*) взяли разность между двумя независимыми средними. В качестве типа анализа мощности (*Type of power analysis*) мы выбрали апостериорный (Post hoc.) В разделе входных параметров (*Input parameters*) мы выбрали двусторонний критерий, ввели оценочный размер эффекта на генеральной совокупности $d = 0.55$, задали частоту ошибок типа I $\alpha = 0.05$ и ввели планируемые размеры выборок $n_1 = n_2 = 40$. Программа строит графики в верхней части окна, а выходные параметры показывает справа внизу. На графике показаны выборочные распределения (рис. 3.7), которые должны быть порождены нулевой (красная кривая) и конкретной альтернативной гипотезой (синяя кривая). Закрашенная синим цветом область обозначена β, чтобы показать, что это частота ошибок типа II. Это вероятность незначимого результата, если $\delta = 0.55$. Мощность равна дополнению к частоте ошибок типа II. Как видим, для заданных входных параметров вычисленная мощность равна 0.68. Это означает, что при заданных условиях значимый результат будет получен с вероятностью 0.68.

Предположим, что мы не удовлетворены вероятностью 0.68 и хотим найти размер выборок, при которых шанс отвергнуть нулевую гипотезу составляет 90 %. В списке *Type of power analysis* (Тип анализа мощности) выберем «A priori», а в модифицированном разделе входных параметров изменим значение поля Power с 0.68 на 0.9. На рис. 7.3 показан вывод программы в новой ситуации. В разделе выходных параметров мы видим, что для получения мощности 0.9 для двустороннего двухвыборочного *t*-критерия, когда размер эффекта на генеральной совокупности $\delta = 0.55$, необходимы размеры выборок $n_1 = n_2 = 71$.

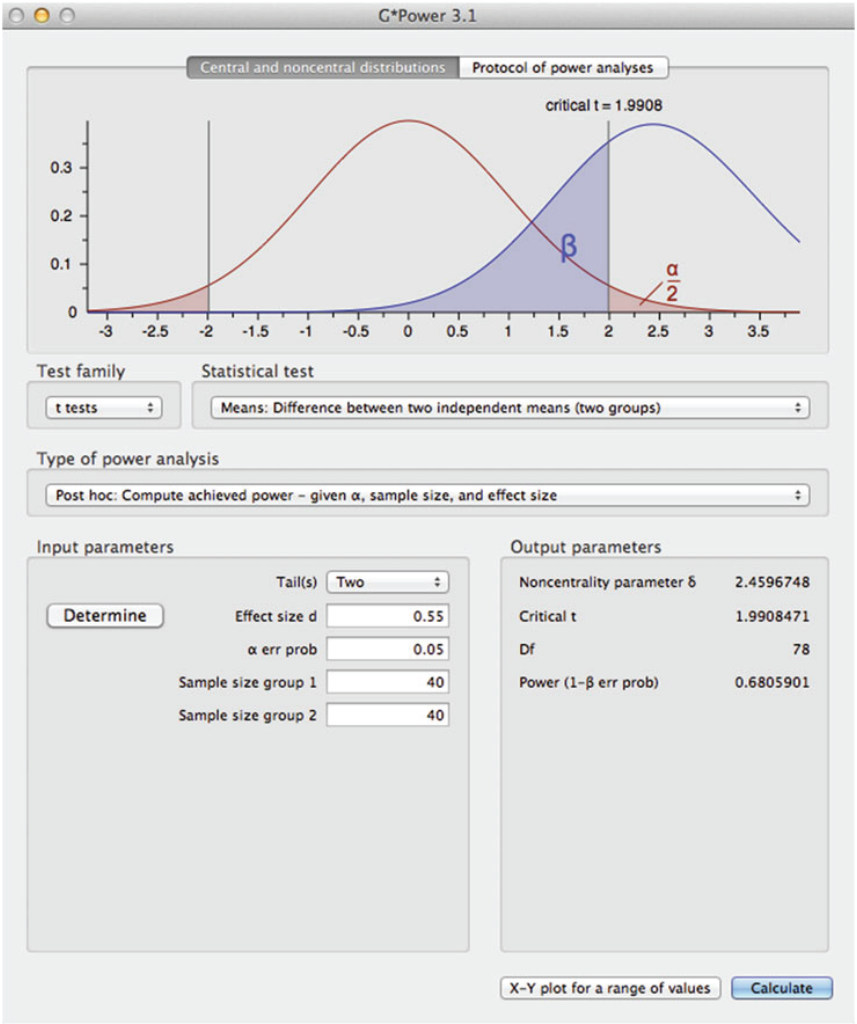


Рис. 7.2. Вывод программы G*Power, вычисляющей мощность *t*-критерия с заданными размерами выборок. В данном случае размер эффекта (0.55) и размеры выборок ($n_1 = n_2 = 40$) известны, а мы ищем мощность, т. е. вероятность получить значимый результат при таких размерах эффекта и выборок и использовании независимого *t*-критерия и $\alpha = 0.05$. Выходными параметрами являются параметр нецентральности δ, который не совпадает с размером эффекта в генеральной совокупности и здесь игнорируется, критическое *t*-значение, число степеней свободы *Df* и самое важное – мощность

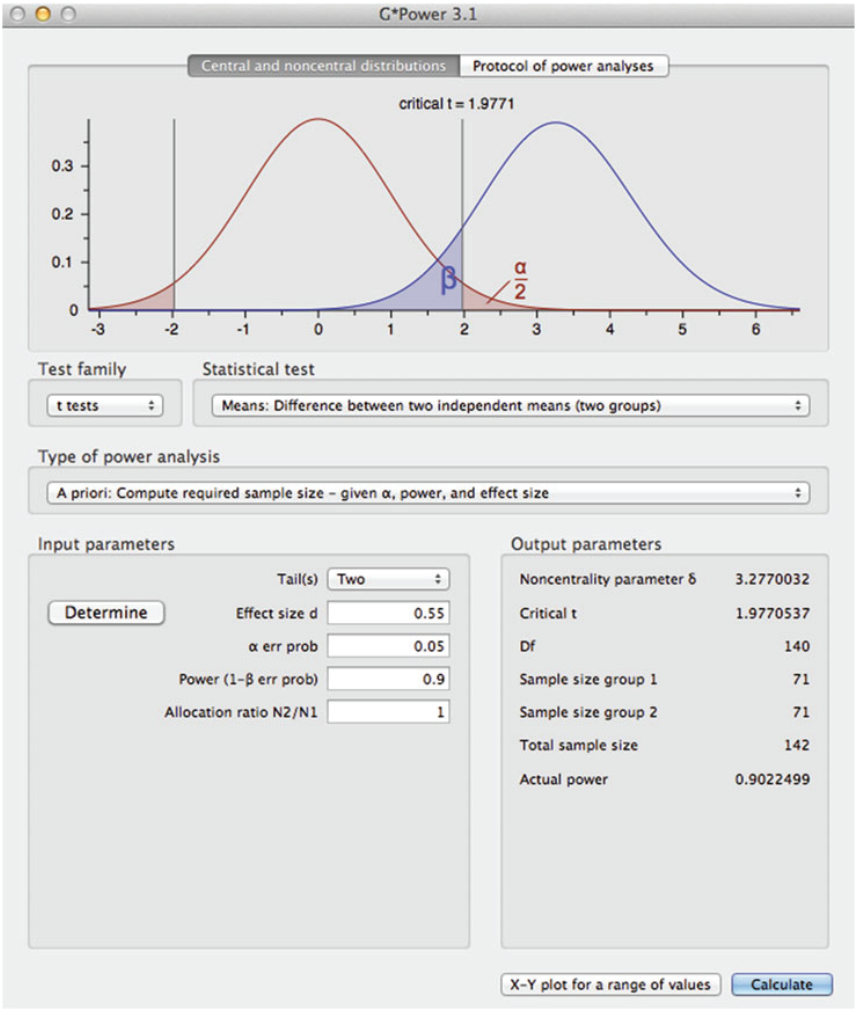


Рис. 7.3. Вывод программы G*Power, вычисляющей размеры выборок, необходимые для того, чтобы мощность *t*-критерия для эксперимента была не менее 90 %. В данном случае размер эффекта (0.55) известен или является желаемым, а ищем мы размер выборки, необходимый для получения значимого результата с вероятностью 0.9.

В общем случае для заданного размера эффекта можно найти, при каких наименьших размерах выборок эксперимент будет иметь заданную мощность. Вычисление таких размеров выборок – важная часть планирования эксперимента. Обычно не имеет особого смыс-

ла ставить эксперимент, не будучи уверенным, что вероятность его успеха, т. е. мощность, достаточна велика. К сожалению, многие ученые ставят эксперименты, не выполнив предварительно анализ мощности, потому что не держат в уме определенный размер эффекта. Такие исследования могут оказаться ценными, но это в значительной мере вопрос удачи. Если вы не можете выполнить осмысленный анализ мощности (с обоснованным размером эффекта), то лучшее, что можно сделать, – «надеяться», что эксперимент даст значимый результат. Если не получилось, то не на кого пенять, потому что у вас никогда и не было (количественных) причин ожидать, что выборка достаточно велика для демонстрации эффекта. Сколько раз ученые занимались предварительной проработкой, полагая, что собираются что-то подтвердить! Подтверждающая работа почти всегда основана на знаниях о размере эффекта, которые можно использовать для планирования эксперимента с высокой мощностью.

7.3. Возможное снижение мощности при сложном плане эксперимента

В идеале анализ мощности производится до сбора данных; это называется априорной мощностью. Но можно также оценивать мощность апостериорно, используя размеры выборок и оцененный на основе данных размер эффекта. В простых случаях (например, двухвыборочный *t*-критерий) апостериорный анализ мощности не говорит ничего, кроме критерия значимости. Если вы воспользуетесь программой G*Power для вычисления мощности при различных комбинациях *t* и размеров выборок, то обнаружите, что если *t*-критерий дает $p > 0.05$, то вычисленная мощность будет меньше 0.5. Аналогично если *t*-критерий дает $p < 0.05$, то вычисленная мощность будет больше 0.5. Если *t*-критерий дает $p = 0.05$, то вычисленная мощность будет приближенно равна 0.5. В этом разделе мы покажем, что апостериорное вычисление мощности может быть полезнее для сложных видов статистического анализа, включающих применение нескольких критериев к набору данных.

Выше мы видели, как использовать G*Power для вычисления мощности при планировании простых экспериментов. Эта и похожие программы вычисляют за раз только один статистический критерий. На практике исследователи часто используют комбинации критериев для обоснования теоретической интерпретации. Для повышения статистической мощности комбинации критериев часто требуется

генерировать модельные наборы данных, соответствующие плану эксперимента и размерам выборок. Такие модельные данные затем анализируются так же, как экспериментальные. Повторив этот процесс тысячи раз, можно просто подсчитать, как часто полный набор статистических выходов соответствует выходам, необходимым для поддержки теоретического предположения. Такой подход на основе моделирования позволяет исследователю рассматривать «вероятность успеха», которая обобщает понятие мощности.

Мы увидим, что у сложных планов эксперимента с несколькими критериями могут быть проблемы с достижением высокой мощности. Даже если мощность отдельных критериев приемлема, может случиться, что у полного набора критериев мощность низкая.

Для демонстрации этого обобщения полезно рассмотреть конкретный пример. Он специально усложнен, потому что запутанность высвечивает важные аспекты анализа мощности. В известном исследовании, опубликованном в 2017 году, приводились эмпирические факты в пользу того, что эффективность запоминания связана с дыханием через нос. Побудительным мотивом для исследования было то, что дыхание через нос может стимулировать мозговую активность в гиппокампе, связанном с обработкой запоминания. Напротив, в дыхании через рот гиппокамп не участвует, поэтому влияния на запоминание быть не должно. Обследуемых просили дышать либо через рот, либо через нос на этапе просмотра изображений во время фазы кодирования информации в памяти. Затем при проведении теста на извлечение из памяти обследуемых просили идентифицировать ранее виденные изображения. В ходе обоих тестов – кодирования и извлечения – изображения предъявлялись в случайные моменты времени – иногда на вдохе, а иногда на выдохе. Главный вывод заключался в том, что верность идентификации была лучше для изображений, предъявленных обследуемым, дышащим через нос на вдохе. И это имело место как при кодировании, так и при извлечении изображений. С другой стороны, дышащие через рот не показали значимого эффекта вдоха по сравнению с выдохом.

Само исследование и анализ его результатов довольно сложны, поэтому полезно охарактеризовать все проверки гипотез. Для удобства перечислим также релевантные статистики из этого исследования. Во всех проверках сравнивалась способность обследуемых к запоминанию.

1. Дышащие через нос ($n_1 = 11$) показали значимый ($F(1, 10) = 6.18$, $p = 0.03$) основной эффект фазы дыхания (вдох или выдох) на способность к запоминанию.

2. Дышащие через нос продемонстрировали лучшую способность к запоминанию изображений, которые извлекались на вдохе, чем на выдохе ($t(10) = 2.85$, $p = 0.017$).
3. Дышащие через рот ($n_2 = 11$) не продемонстрировали лучшую способность к запоминанию изображений, которые извлекались на вдохе, чем на выдохе ($t(10) = -1.07$, $p = 0.31$).
4. В целом не было зарегистрировано значимого различия между дышащими носом и ртом ($F(1, 20) = 1.15$, $p = 0.29$).
5. Имело место значимое взаимодействие между фазой дыхания (вдох или выдох) и типом дыхания (через нос или через рот), когда изображения помечались способом кодирования (на вдохе или на выдохе) ($F(1, 20) = 4.51$, $p = 0.046$).
6. Также имело место значимое взаимодействие между фазой дыхания (вдох или выдох) и типом дыхания (через нос или через рот), когда изображения помечались способом извлечения (на вдохе или на выдохе) ($F(1, 20) = 7.06$, $p = 0.015$).

Если вы запутались, то черпайте утешение в том, что вы не одиноки. Исследование и анализ его результатов очень сложны, поэтому читателю трудно связать опубликованные статистики с теоретическими выводами. К тому же некоторые сравнения кажутся неуместными. Например, авторы исследования использовали проверки 2 и 3 для демонстрации различий в значимости между дышащими через нос и через рот (сравнивая качество извлечения из памяти на вдохе и на выдохе). В главе 3 (следствие 3b) мы отмечали, что различная значимость не то же самое, что значимое различие. Аналогично авторы сочли, что нулевой результат в проверке 3 указывает на «отсутствие различий» в общей эффективности запоминания при дыхании через нос и через рот. В главе 2 (следствие 3a) мы видели, что отсутствие доказательства не то же самое, что доказательство отсутствия.

Таблица 7.1. Оценки вероятностей успеха для находок, обнаруженных в исследовании, связывающем эффективность запоминания с типом дыхания (через нос или через рот)

Проверка	Вероятность успеха
Через нос: основной эффект фазы дыхания	0.690
Извлечение при дыхании через нос: эффект фазы дыхания	0.655
Извлечение при дыхании через рот: эффект фазы дыхания	0.809

Окончание табл. 7.1

Проверка	Вероятность успеха
Сравнение дыхания через нос и через рот: нулевой основной эффект	0.820
На этапе кодирования: взаимодействие между фазой и типом дыхания	0.604
На этапе извлечения: взаимодействие между фазой и типом дыхания	0.708
Все проверки	0.216

Оставим на время наши сомнения относительно уместности проверок. Для успеха этого исследования требовалось получить четыре значимых результата и два незначимых. Если бы какой-то из этих результатов оказался безуспешным, то были бы поставлены под сомнение некоторые выводы авторов. Как оказалось, данные подтвердили каждый из необходимых результатов. Мы покажем, что при таком большом числе результатов, которые должны быть подтверждены одним набором данных, столь полный успех был бы редкостью, даже если бы эффекты действительно имели место и были близки к оценкам значений, полученных на экспериментальных данных. Чтобы оценить вероятность такого успеха, мы воспользовались статистической программой R для генерирования 100 000 модельных экспериментов с такими же размерами выборки, средними, стандартными отклонениями и корреляциями (внутрисубъектных аспектов эксперимента), как указано в исследовании. В табл. 7.1 показано, как часто каждая проверка давала желаемый результат. Вероятность успеха для любой отдельно взятой гипотезы варьируется от 0.60 до 0.82. Для каждой значимой проверки вероятность ее успеха соответствует мощности. Для проверок 3 и 4 успешным считался незначимый результат, и в таблице приведена вероятность *не* отвергнуть нулевую гипотезу.

Однако вероятность того, что при таком моделировании *каждая* проверка будет успешна, гораздо ниже вероятности успеха одной проверки, потому что данные должны обладать свойствами, необходимыми для получения значимого результата в одних проверках и незначимого в других. Совместная вероятность успеха всех проверок в одном эксперименте составляет всего 0.216. Такое низкое значение наводит на мысль, что на случайной выборке прямое воспроизведение этого исследования с похожим размером выборки имело бы довольно низкие шансы повторить описанные результаты.

Исследователь, задумавший повторить это исследование, наверно, захотел бы взять выборку такого размера, при котором вероятность успеха высока. Увеличение выборки повышает мощность критерия, поэтому исследование всего с одной проверкой имело бы больше шансов обнаружить эффект, если он существует. Но когда теоретические выводы основываются как на значимых, так и на незначимых проверках, максимальная вероятность успеха ограничена, потому что при увеличении размера выборки малые эффекты порождают значимые результаты (даже если авторы исследования надеются на нулевой результат). Предел вероятности для этого исследования можно изучить с помощью дополнительных модельных экспериментов, в которых размер выборки варьируется для каждого условия. Цветные линии на рис. 7.4 показывают оценочные вероятности успеха для каждой из шести проверок в виде функции от размера выборки (в предположении, что размер выборки одинаков для каждого условия). Для тех четырех проверок, в которых успехом считается получение значимого результата, вероятность успеха возрастает вместе с ростом размера выборки и сходится к максимальному значению 1 при размере выборки около 40. Для двух проверок, где успехом считается незначимый результат, вероятность успеха убывает с ростом размера выборки (поскольку в некоторых случайных выборках имеют место значимые различия). Штриховые черные линии на рис. 7.4 показывают, что при рассмотрении всех шести проверок вместе (черная линия) максимально возможная вероятность успеха составляет 0.37, когда выбрано $n_1 = n_2 = 20$ обследуемых для каждого условия.

Этот анализ вероятности успеха наводит на мысль, что для исследования связи между типом дыхания и эффективностью запоминания желателен другой план эксперимента. Простые эксперименты обычно лучше, потому что чем больше требований предъявляется к набору данных (например, порождать много значимых или незначимых результатов), тем меньше вероятность того, что какой-то конкретный набор данных даст требуемый набор результатов. Принимая во внимание низкую оценочную вероятность успеха в этом исследовании, возникает вопрос, как авторам удалось так удачно сформировать случайную выборку, что были подтверждены и отвернуты именно те гипотезы, которые были им необходимы для подтверждения своих теоретических предположений. Мы рассмотрим этот вопрос в главе 10 при обсуждении того, как следует интерпретировать статистику в повторных экспериментах.

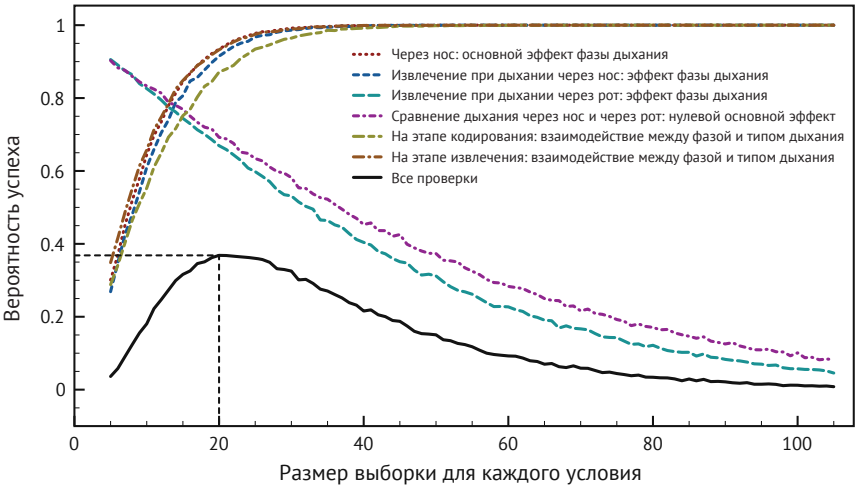


Рис. 7.4. Каждая цветная линия показывает оценочную вероятность успеха в виде функции от размера выборки для одной проверки из исследования влияния типа дыхания на эффективность запоминания. Сплошная черная кривая показывает оценочную вероятность успеха всех проверок. Штриховые черные линии маркируют размер выборки, для которой вероятность успеха сразу всех проверок максимально возможная. Каждое значение основано на 10 000 модельных экспериментов

Что следует запомнить

1. Старайтесь составлять простые планы: подумайте об уплотнении исходных данных в промежуточные переменные, которые уже подвергаются статистическому анализу.
2. Производите анализ мощности до начала эксперимента, чтобы оценить, есть ли у него реальные шансы продемонстрировать существующий эффект.
3. Старайтесь составлять простые планы: если для подтверждения теории необходимы как значимые, так и нулевые результаты, то мощность может сильно снизиться.

Глава 8

Корреляция

Что вы узнаете из этой главы

В главах 3 и 6 мы изучали влияние географической широты на высоту деревьев, для чего обмеряли деревья в двух и трех местах и проверяли, различаются ли средние высоты. Как мы увидим ниже, для ответа на этот вопрос лучше измерять высоту на большем числе параллелей. Применение дисперсионного анализа (ANOVA) не лучшее решение в этой ситуации, потому что не принимается во внимание тот факт, что широта измеряется по относительной шкале. В ANOVA все широты трактуются как номинальные значения (см. главу 7). Корреляция позволяет учесть относительность шкалы и тем самым выразить влияние широты одним значением, *r*.

8.1. Ковариация и корреляция

Сначала поговорим о визуализации корреляции. В случае отрицательной линейной корреляции увеличению одной переменной соответствует линейное уменьшение другой переменной, например высота деревьев уменьшается с возрастанием широты. Если отложить широту по оси *x*, а высоту деревьев по оси *y*, то точки лягут на прямую, как показано на рис. 8.1a (отрицательная линейная корреляция). С другой стороны, если две переменные никак не связаны между собой, то данные будут выглядеть как размытое облако точек (рис. 8.1c, корреляция отсутствует). Если высота деревьев возрастает с увеличение широты, то имеет место положительная линейная корреляция (рис. 8.1e). Обычно мы имеем что-то среднее (рис. 8.1b, d).

Рис. 8.1. Пять случаев зависимости высоты деревьев от географической широты. Каждая точка представляет высоту одного дерева в одном месте. Корреляция измеряет степень линейной зависимости между двумя переменными

Эта линейная связь описывается формулой ковариации:

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X}) \times (y_i - \bar{Y})}{n-1}, \tag{8.1}$$

где x_i – например, значения широты, y_i – значения высоты деревьев, а X и Y – средние, т. е. средняя широта и средняя высота дерева соответственно. Данные представляют собой n пар (широта, высота дерева). Ковариация обобщает понятие дисперсии, потому что cov(x, x) – дисперсия x.

Недостаток ковариации в том, что она зависит от масштаба. Например, если высота деревьев измеряется в метрах, то ковариация будет меньше, чем при измерении в сантиметрах. Поэтому мы нормируем ковариацию, деля ее на произведение стандартных отклонений x и y, и таким образом приходим к корреляции:

$$r = \frac{\text{cov}(x, y)}{s_x s_y}. \tag{8.2}$$

Корреляция этого типа называется коэффициентом корреляции Пирсона. Его значения изменяются от −1.0 до +1.0, где −1 означает отрицательную линейную корреляцию, 0 – отсутствие корреляции, а +1 – положительную линейную корреляцию (рис. 8.1a, c и e). Другие значения соответствуют промежуточной тесноте связи между переменными.

8.2. Проверка гипотез с помощью корреляции

На рис. 8.2 приведен пример (n = 50) данных о высоте деревьев на разных широтах. Каждая точка соответствует одному дереву. Очевидно, что линейной корреляции нет, но кажется, что корреляция все

же отлична от нуля. Мы воспользуемся проверкой гипотез, чтобы узнать, имеется ли значимая корреляция. Нулевая гипотеза имеет вид:

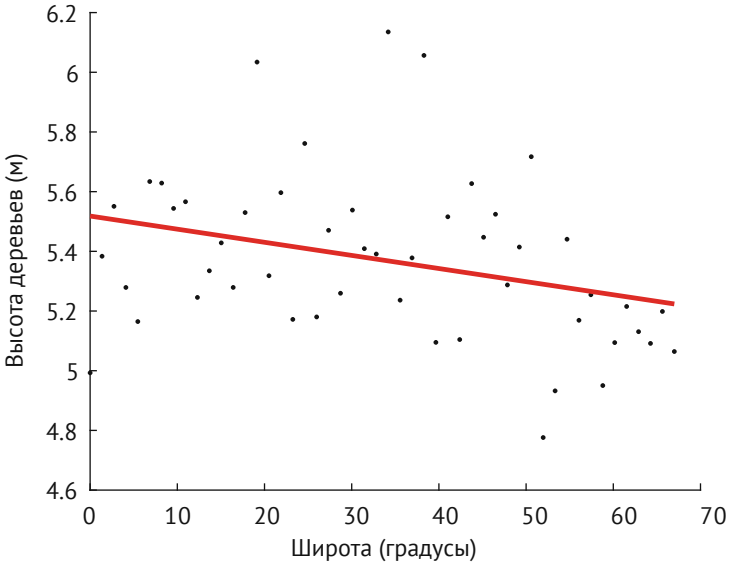


Рис. 8.2. Зависимость высоты деревьев от географической широты для выборки из 50 деревьев. Коэффициент корреляции равен $r = -0.312$. Красная линия – наилучшая эмпирическая прямая

$$H_0 : \rho = 0,$$

где ρ – корреляция в генеральной совокупности.

Мы не будем вдаваться в детали, но если нулевая гипотеза верна, то стандартное отклонение выборочного распределения выборочной корреляции равно:

$$s_r = \sqrt{\frac{1-r^2}{n-2}}, \tag{8.3}$$

а соответствующая статистика критерия – t-значение, вычисляемое по формуле:

$$t = \frac{r-0}{s_r} \tag{8.4}$$

с числом степеней свободы $df = n - 2$. Вывод типичной статистической программы для данных на рис. 8.2 выглядит, как показано

в табл. 8.1.

Таблица 8.1. Вывод корреляции типичной статистической программой

r	*t*	*df*	*p*
−0.312	−2.28	48	0.027

Поскольку значение *p* меньше 0.05, мы заключаем, что имеется значимая корреляция. Тот факт, что значение *r* отрицательно, означает, что более высокие деревья растут на меньшей широте.

8.3. Интерпретация корреляции

Предположим, что мы выявили значимую корреляцию между переменными *x* и *y*. О чем это говорит? Во-первых, это не значит, что *x* является причиной *y*. Это легко понять, заметив, что

$$\mathrm{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{X})\times(y_i-\bar{Y})}{n-1}=\frac{\sum_{i=1}^{n}(y_i-\bar{Y})\times(x_i-\bar{X})}{n-1}=\mathrm{cov}(y,x), \qquad (8.5)$$

что при неправильной интерпретации означало бы, что *x* является причиной *y* и *y* является причиной *x*. Значимая корреляция может иметь место по четырем причинам:

1) *x* является причиной *y*,
2) *y* является причиной *x*,
3) некоторая промежуточная переменная *z* является причиной *x* и *y*,
4) корреляция ложная.

Приведем пример промежуточной переменной (причина 3): не сама широта определяет высоту деревьев, а некоторые факторы, связанные с широтой, например запас воды. Ложная корреляция (причина 4) может возникнуть случайно. Например, за период с 2000 по 2009 год между потреблением сыра на душу населения в США и количеством людей, умерших, потому что запутались в простыне, имеется корреляция *r* = 0.947. Если бы ученый обнаружил такую высокую корреляцию в эксперименте, то откупорил бы бутылку шампанского! Ложные корреляции неизбежны, если рассматривать достаточно много наборов данных.

Важно отметить, что корреляция измеряет только линейную связь, поэтому незначимая корреляция еще не означает, что между *x* и *y* нет никакой причинно-следственной связи. Например, темпе-

ратура воздуха систематически изменяется на протяжении суток по синусоидальному закону (поднимается днем и опускается ночью), но коэффициент корреляции между временем суток и температурой $r \approx 0$.

Всегда рекомендуется не только вычислять коэффициент корреляции, но и изучать график данных. Данные совершенно разных типов могут приводить к одному и тому же значению r (рис. 8.3), поэтому знание одной лишь корреляции несет только частичную информацию о наборе данных. Ко всему прочему, корреляция очень чувствительна к выбросам (рис. 8.4), и добавление или удаление всего одной точки может кардинально изменить коэффициент корреляции набора данных.

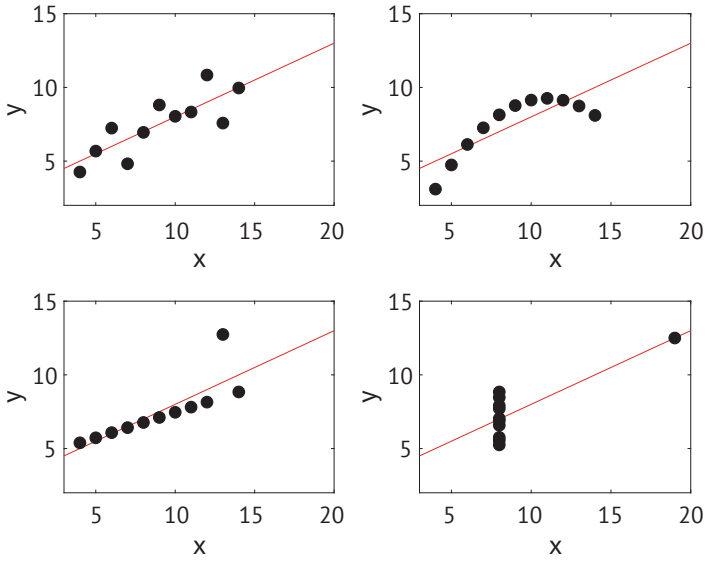


Рис. 8.3. Квартет Энскомба. Для всех наборов данных значение r одно и то же ($r = 0.816$), хотя на графике они ничуть не похожи

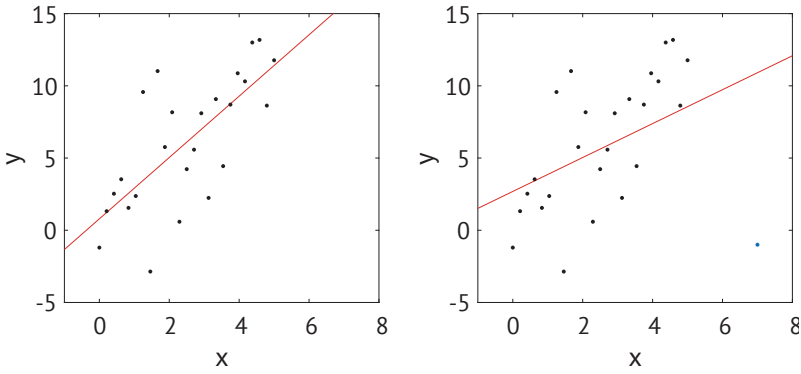


Рис. 8.4. Выбросы могут оказать сильное влияние на корреляцию. Слева: оригинальные данные с $r = 0.71$. Справа: добавление единственного выброса (синяя точка в правом нижнем углу) привело к резкому уменьшению коэффициента корреляции ($r = 0.44$)

Таблица 8.2. Рекомендации Коэна по размеру эффекта для |r|

	Малый	Средний	Большой
Размер эффекта	0.1	0.3	0.5

8.4. Размер эффекта

Корреляцию часто применяют как меру размера эффекта, показывающую тесноту связи между двумя переменными. В частности, квадрат корреляции, r^2, показывает, какая доля изменчивости одной переменной (например, высоты деревьев) может быть объяснена изменчивостью другой переменной (например, широты). Это информация того же рода, что и величина η^2, рассмотренная в главе 6. Согласно Коэну значение r, меньшее 0.1, считается малым эффектом, как и значение, большее −0.1 (табл. 8.2).

8.5. Сравнение с подгонкой модели, ANOVA и *t*-критерием

В главе 7 мы подгоняли модель к данным об обучении и обращали внимание прежде всего на угловой коэффициент, который аналогичен коэффициенту корреляции, потому что корреляция – это мера

линейности связи. Проверка гипотезы о ненулевом угловом коэффициенте дает такой же результат, как проверка гипотезы о ненулевой корреляции.

Как было отмечено в главе 7, не стоит использовать ANOVA, когда независимая переменная измеряется по относительной шкале, потому что в ANOVA считается, что независимая переменная номинальная. Поскольку анализ, основанный на корреляции, задействует все преимущества относительной шкалы, его мощность выше, чем у ANOVA.

Можно было бы также использовать *t*-критерий, разбив данные на две части, например больше и меньше медианной широты, т. е. половину данных отнести к северной группе, а другую половину – к южной. В общем случае такой подход хуже анализа, основанного на корреляции, потому что (снова) не учитывает относительность шкалы независимой переменной. Например, на рис. 8.5 данные, показанные на рис. 8.2, разделены между регионами с высокими и низкими широтами. *t*-критерий не дает значимого результата. Таким образом, если проанализировать данные с таким разбиением на подмножества, то мы не заметим значимого различия, которое нашли в результате анализа корреляции исходного набора данных (табл. 8.1).

В некотором смысле корреляцию можно рассматривать как обобщение дисперсионного анализа и *t*-критерия.

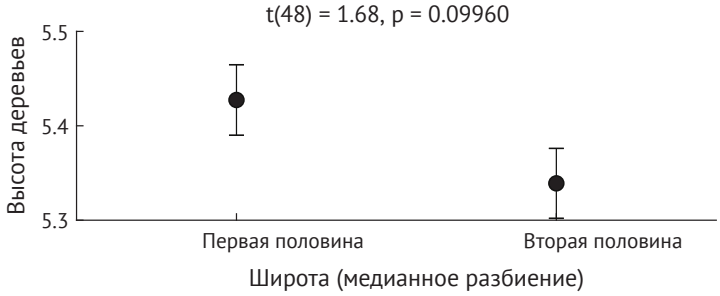


Рис. 8.5. Данные, порожденные в результате медианного разбиения данных на рис. 8.2. Исследование с помощью *t*-критерия не показывает значимого различия между средними

8.6. Предположения и подводные камни

Для проверки гипотез о наличии корреляции необходимо сделать несколько предположений.

1. Как всегда, данные должны быть независимы и одинаково распределены.

2. Распределение переменной *y*, обусловленной любым значением переменной *x*, должно быть нормальным. То есть если мы возьмем все значения *y* при одном и том же значении *x* и построим их гистограмму, то эта гистограмма будет иметь нормальное распределение.
3. Шкала обеих переменных интервальная или относительная.
4. Размер выборки фиксируется до начала эксперимента.

Если шкала данных порядковая, то можно вычислить коэффициент корреляции Спирмена, в котором используются ранги (порядковая шкала), а относительная шкала необязательна. Коэффициент корреляции Спирмена – непараметрический эквивалент параметрического коэффициента корреляции Пирсона.

8.7. Регрессия

В этом подразделе мы кратко опишем связь между корреляцией и регрессией. Если торопитесь, можете пропустить его. В следующих главах регрессия не будет играть никакой роли.

Корреляция говорит, насколько плотно упакованы данные вокруг оптимальной эмпирической прямой. Например, коэффициент корреляции 1.0 означает, что данные точно ложатся на прямую. Но что такое эта оптимальная эмпирическая прямая? Регрессия дает ее уравнение с двумя параметрами: угловой коэффициент *m* и свободный член *b* (ордината точки пересечения прямой с осью *y*). Угловой коэффициент прямой регрессии равен стандартному отклонению в направлении *y*, поделенному на стандартное отклонение в направлении *x* и умноженному на весовой коэффициент *r* из формулы 8.2:

$$m = r\frac{s_y}{s_x}. \tag{8.6}$$

Таблица 8.3. Вывод регрессии типичной статистической программой

Параметр	Значение коэффициента	*t*	*p*
Свободный член (константа)	12.146	4.079	0.00017
Угловой коэффициент (широта)	–0.147	–2.275	0.027

Это означает, что для любой величины стандартного отклонения, пройденной в направлении x, мы поднимаемся в направлении y на величину стандартного отклонения, умноженную на r.

Свободный член b равен:

$$b = \bar{y} - m\bar{x}.$$

Для данных о высоте деревьев на рис. 8.2 угловой коэффициент $m = -0.1473$, а свободный член $b = 12.1461$. Это означает, что на широте 0° средняя высота деревьев равна 12.1461 м и что на каждый градус широты к северу от экватора высота деревьев изменяется на −0.1473 м (т. е. при увеличении широты высота деревьев уменьшается). Типичная статистическая программа показывает эти результаты в виде таблицы, как в табл. 8.3.

Здесь, помимо углового коэффициента и свободного члена прямой регрессии, программа выводит также t- и p-значение того и другого – так называемые коэффициенты регрессии.

Эти статистики проверяют нулевую гипотезу, согласно которой угловой коэффициент и свободный член равны нулю. В данном примере p-значение меньше 0.05, поэтому обе величины значимо отличны от нуля. В такой ситуации соответствующий коэффициент корреляции (значение r) обычно значимо отличается от нуля. Если свободный член значимо не отличается от нуля, то прямая регрессии приближенно проходит через точку (0, 0).

Что следует запомнить

1. Корреляцию следует предпочесть, если обе переменные измеряются по интервальной или относительной шкале.
2. Не следует путать причинно-следственную связь с корреляцией.
3. Совершенно различные наборы данных могут приводить к одному и тому же значению r.

Часть III
Метаанализ и кризис науки

Глава 9

Метаанализ

Что вы узнаете из этой главы

В части III мы покажем, что путем комбинирования данных из различных экспериментов можно получить совершенно новые знания. Например, даже если статистический результат каждого эксперимента сам по себе не имеет особого смысла, иногда комбинация данных указывает на наличие проблемы. Насколько вероятно, что все четыре эксперимента с малым эффектом и малым размером выборки приведут к значимым результатам? Мы покажем, что чаще всего крайне маловероятно. Отсюда вытекает простое следствие: если эксперименты всегда дают значимые результаты, то данные слишком хороши, чтобы быть правдой. Мы покажем, как распространенная, но неправильно понятая научная практика ведет к «слишком хорошим, чтобы быть правдой» данным, как эта практика вздувает частоту ошибок типа I и как это привело к серьезному кризису науки, затронувшему многие области, где статистика играет ключевую роль. В этом смысле часть III можно считать обобщением следствий, сформулированных в главе 3. И в конце мы обсудим потенциальные решения.

В этой главе мы обобщим стандартизованные размеры эффектов из главы 2 и покажем, как комбинировать данные из разных экспериментов для вычисления метастатистики.

9.1. Стандартизованные размеры эффектов

Как было сказано в части I, статистика в значительной своей части занимается различением зашумленного сигнала и чистого шума. Для стандартного двухвыборочного *t*-критерия отношение сигнала к шуму называется *d* Коэна и оценивается на основе данных по следующей формуле (см. главу 3):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

d Коэна говорит нам, насколько просто различить два средних. Разность средних находится в числителе. Большую разность проще обнаружить, чем малую, но нужно также принимать во внимание шум. При большом стандартном отклонении обнаружить различие между средними труднее (см. главу 2). Если $n_1 = n_2 = n$, то *t*-значение в двухвыборочном *t*-критерии равно:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{2}{n}}} = \frac{d}{\sqrt{\frac{2}{n}}} = d\sqrt{\frac{n}{2}}.$$

Таким образом, *t*-значение является просто произведением *d* Коэна на вес, равный функции от размера выборки. Как отмечалось в главе 3, всегда полезно проверять размер эффекта. К сожалению, во многих исследованиях указывается лишь *p*-значение, в котором размер эффекта и размер выборки объединены. Исходя из приведенной выше формулы, мы можем вычислить *d* Коэна, зная *t*-значение и размер выборки:

$$d = t\sqrt{\frac{2}{n}}.$$

Важное свойство *d* Коэна – независимость его абсолютной величины от размера выборки. Это следует из того, что *d* – оценка фиксированного (хотя и неизвестного) значения для генеральной совокупности[1].

В главе 3 мы показали, что δ можно оценить с помощью *d*. Однако *d* является хорошей оценкой, только если выборка велика. Для относительно небольших выборок *d* систематически завышает оценку размера эффекта в генеральной совокупности δ. Это завышение можно скорректировать, воспользовавшись не *d*, а *g* Хеджеса:

$$g = \left(1 - \frac{3}{4(2n-2)-1}\right)d.$$

[1] Отметим, что, хотя в эту конкретную формулу размер выборки *n* входит, он лишь компенсирует увеличение *t* с ростом размера выборки.

Почти для всех практических целей можно считать, что *g* Хеджеса – то же самое, что *d* Коэна. Мы упомянули его здесь только потому, что будем использовать для вычислений в контексте метаанализа. В приложение к этой главе включены формулы для случая $n_1 \neq n_2$ и других типов планирования эксперимента.

9.2. Метаанализ

Предположим, что один и тот же (или очень похожие) эксперимент выполняется несколько раз. На первый взгляд, ничто не мешает объединить данные нескольких экспериментов, чтобы прийти к еще более убедительным выводам и достичь большей мощности. У такого объединения есть даже специальное название – метаанализ. Оказывается, что для проведения такого метаанализа очень полезны стандартизованные размеры эффектов.

В табл. 9.1 сведены статистические показатели из пяти исследований, в которых авторы приходят к выводу, что обращение с деньгами уменьшает нищету по сравнению с социальной изоляцией. Во всех исследованиях использовался двухвыборочный *t*-критерий, а в столбце *g* приведено значение *g* Хеджеса, являющееся просто оценкой размера эффекта.

Чтобы объединить размеры эффектов по всем исследованиям, необходимо принять во внимание размеры выборок. Эксперимент с участием 46 обследуемых весит больше, чем эксперимент с 36 обследуемыми в каждой группе. В последнем столбце табл. 9.1 приведен взвешенный размер эффекта, $w \times g$, для каждого эксперимента (о вычислении *w* см. приложение). Объединенный размер эффекта вычисляется путем сложения взвешенных размеров эффектов и деления результата на сумму весов:

$$g^* = \frac{\sum_{i=1}^{5} w_i g_i}{w_i} = 0.632.$$

Этот метааналитический размер эффекта – лучшая оценка, основанная на результатах пяти экспериментов. Стоит ли объединять стандартизованные размеры эффектов таким образом, сильно зависит от теоретической интерпретации эффектов. Если теория говорит, что все эксперименты измеряют, по существу, один и тот же эффект, то такой вид объединения допустим, и результатом станет более точная оценка размера эффекта. С другой стороны, нет особо-

го смысла объединять существенно различные эксперименты, в которых измеряются разные эффекты.

Метаанализ сильно осложняется, если эксперименты структурно различны (например, в опубликованных работах могут применяться *t*-критерии, ANOVA или корреляция). Но, несмотря на трудности, метаанализ может стать удобным средством объединения экспериментальных данных и, следовательно, получения более точных оценок эффекта.

Таблица 9.1. Данные из пяти экспериментов, использованные для метаанализа

n	*t*	*g*	*w* × *g*
36	3.01	0.702	12.15
36	2.08	0.485	8.66
36	2.54	0.592	10.43
46	3.08	0.637	14.17
46	3.49	0.722	15.83

Что следует запомнить

1. Объединение размеров эффектов из разных экспериментов дает более точные оценки.
2. Объединение экспериментальных данных повышает мощность.

Приложение

Стандартизованные размеры эффектов в более сложных случаях

Если размеры выборок различны ($n_1 \neq n_2$), то *t*-значение двухвыборочного *t*-критерия равно:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{d}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}}.$$

Если в опубликованном исследовании не приводятся средние и стандартные отклонения выборок, то *d* Коэна можно вычислить по указанным *t*-значению и размеру выборки:

$$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}.$$

g Хеджеса при неравных размерах выборок вычисляется по формуле:

$$g = \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right)d.$$

Существуют аналогичные формулы стандартизованных размеров эффектов и поправок для других планов эксперимента. Например, для одновыборочного *t*-критерия с нулевой гипотезой, заключающейся в том, что среднее генеральной совокупности равно *a*, величина *d* Коэна вычисляется по формуле:

$$d = \frac{\bar{x} - a}{s},$$

где сигнал (отклонение от значения, предполагаемого в нулевой гипотезе), как и раньше, находится в числителе, а шум (выборочное стандартное отклонение) – в знаменателе. Несмещенным вариантом *d* Коэна для одновыборочного случая является *g* Хеджеса:

$$g = \left(1 - \frac{3}{4(n-1) - 1}\right)d.$$

Для *t*-критериев с повторными измерениями подходящий стандартизованный размер эффекта зависит от способа его использования. Иногда исследователь хочет получить размер эффекта относительно разностных примеров, вычисляемых для каждого обследуемого. В таком случае подходит одновыборочное *d* или *g*. А иногда требуется найти размер эффекта, эквивалентный тому, что был бы получен в двухвыборочном независимом *t*-критерии. Тогда необходимо компенсировать корреляцию между примерами. Если опубликовано *t*-значение зависимой выборки, то вычисления производятся по формуле:

$$d = \frac{t}{\sqrt{n}}\sqrt{2(1-r)}.$$

К сожалению, в большинстве работ не публикуется корреляция между примерами для зависимой выборки. Для наших целей основная идея стандартизованного размера эффекта важнее конкретного вычисления. Однако следует иметь в виду, что формулы, встречающиеся в интернете, иногда подразумевают не высказанные явно предположения, например, о равенстве размеров выборок для независимого *t*-критерия или что $r = 0.5$ для зависимого *t*-критерия.

Более полный пример метаанализа

В табл. 9.2 приведены некоторые промежуточные члены, отсутствующие в табл. 9.1.

Таблица 9.2. Подробный метаанализ с дополнительными вычислениями для тех же данных, что в табл. 9.1

n_1	n_2	t	g	v_g	w	wg
36	36	3.01	0.702	0.058	17.3	12.15
36	36	2.08	0.485	0.056	17.9	8.66
36	36	2.54	0.592	0.057	17.6	10.43
46	46	3.08	0.637	0.045	22.2	14.17
46	46	3.49	0.722	0.046	21.9	15.83

Для объединения размеров эффекта из разных исследований мы умножаем каждое значение *g* на вес, равный его обратной дисперсии. Вычисление обратной дисперсии состоит из нескольких шагов. Для независимого двухвыборочного *t*-критерия формула дисперсии *d* Коэна имеет вид:

$$v_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)},$$

а дисперсия *g* Хеджеса включает квадрат встречавшегося ранее поправочного члена:

$$v_g = \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right)^2,$$

который показан в отдельном столбце табл. 9.2. Для выполнения метаанализа каждый стандартизованный размер эффекта умножается на свою обратную дисперсию:

$$w = \frac{1}{v_g},$$

которая находится в столбце слева от произведения *wg*. Объединенный размер эффекта вычисляется путем суммирования произведений и деления результата на сумму весов:

$$g^* = \frac{\sum_{i=1}^{5} w_i g_i}{\sum_{i=1}^{5} w_i} = 0.632.$$

Глава 10

Воспроизводимость

Что вы узнаете из этой главы

В этой главе используются методы анализа мощности из главы 7 и методы метаанализа из главы 9, чтобы выявить примеры некорректного статистического анализа в опубликованных работах. Основная идея проста. Если мощность не очень велика, то мы знаем, что даже реально существующие эффекты не всегда дают значимые результаты из-за недостаточности случайной выборки. Если при умеренной мощности все результаты получились значимыми, то опубликованные цифры слишком хороши, чтобы быть правдой. Из наших рассуждений вытекает важнейшее следствие: при использовании статистической проверки гипотез воспроизведение не может являться окончательным арбитром в науке, если только экспериментальная мощность не очень велика. В главе 11 показано, что такие результаты могут получаться, даже когда ученый старается все делать правильно.

10.1. Кризис воспроизводимости

Во всех науках воспроизводимость считается «золотым стандартом» для демонстрации важных открытий. Если коллега усомнился в достоверности вашего эмпирического утверждения, то самый верный способ поставить его на место – продемонстрировать, что эффект устойчиво воспроизводится. Такая демонстрация особенно убедительна, если эффект воспроизведен в независимой лаборатории. И наоборот, если независимая лаборатория сообщает, что воспроизвести эффект не удалось, то, скорее всего, последует оживленная дискуссия на тему правильного применения процедур и интерпретации результатов. Успешное воспроизведение высоко ценится и считается сильным аргументом в пользу научного утверждения.

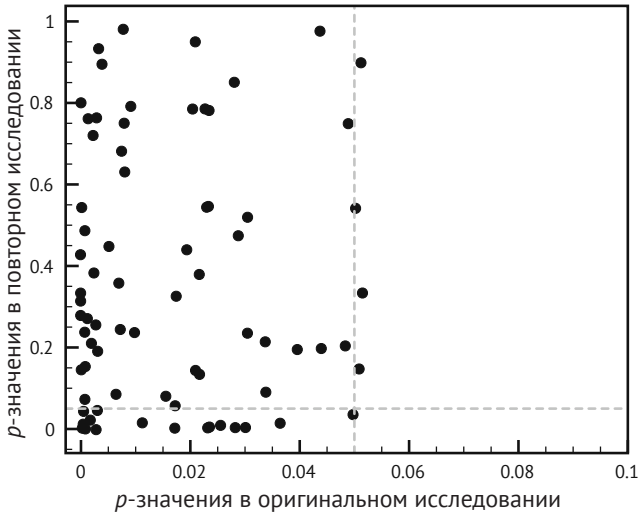


Рис. 10.1. Каждая точка соответствует паре *p*-значений из оригинального исследования и его воспроизведения. Хотя почти во всех оригинальных исследованиях сообщалось, что $p < 0.05$, лишь в очень немногих воспроизведениях удалось получить столь же малое *p*-значение. Рисунок печатается с разрешения Open Science Collaboration [1]. Обратите внимание на сильно различающийся масштаб по осям *x* и *y*. Диапазон значений по оси *x* – от 0.0 до 0.1, а по оси *y* – от 0.0 до 1.0. Не просматривается очевидная связь между *p*-значениями в оригинальном и повторном исследованиях. Хорошим считался бы результат, при котором *p*-значения в повторном исследовании меньше 0.05, т. е. когда все черные точки расположены ниже штриховой горизонтальной прямой

К сожалению, во многих научных дисциплинах с воспроизводимостью дело обстоит неважно. Группа психологов под названием Open Science Collaboration [1] предприняла попытку воспроизвести 97 исследований, опубликованных в трех ведущих журналах. В их отчете от 2015 года лишь 36 % повторных исследований дали результаты, сходные с оригинальными. Каждая точка на рис. 10.1 представляет пару *p*-значений: из оригинального и повторного исследования. Штриховая вертикальная прямая обозначает порог 0.05 в оригинальных исследованиях, и, как видно, почти во всех оригинальных исследованиях *p*-значение ниже этого порога. Оно и не удивительно, потому что обычно публикуются только значимые результаты. Штриховая горизонтальная прямая обозначает порог 0.05 в повторных исследованиях, и почти все *p*-значения оказались выше этого порога. Шокирующим стал тот факт, что между оригиналь-

ным и повторным *p*-значением не прослеживается никакой связи. Например, в некоторых оригинальных исследованиях *p*-значения были гораздо меньше 0.01, а в повторных оказались близки к 1.0. Хуже того, во многих повторных исследования выборка была *больше*, чем в оригинальных, поэтому *p*-значения должны были бы быть меньше, как было подчеркнуто в главе 3 (следствие 2d).

Проблемы воспроизводимости свойственны не только психологии. В 2012 году исследователи из биотехнологической компании Amgen сообщили, что не смогли воспроизвести результаты 47 из 53 считающихся знаковыми статей, относящихся к исследованиям рака. В академических кругах наблюдается стремление повторить то, что было сделано психологами. И первые результаты показывают, что с дела с воспроизводимостью обстоят ничуть не лучше, чем в психологии. Для многих невоспроизводимость результатов в этих исследованиях служит признаком чрезвычайно серьезных проблем, которые иногда называют «кризисом воспроизводимости».

Мы согласны, что проблемы серьезны. Однако думаем, что, вместо того чтобы недоумевать, почему повторные исследования не приносят успеха, проще взглянуть на оригинальные опубликованные результаты и показать, что они никогда и не имели смысла.

Рассмотрим следующие два явления, которые изучались в нескольких экспериментах.

- Явление A: 9 из 10 экспериментов дали значимые результаты, поэтому коэффициент воспроизводимости равен 0.9.
- Явление B: 10 из 19 экспериментов дали значимые результаты, поэтому коэффициент воспроизводимости равен 0.53.

Если согласиться с тем, что успешное воспроизведение – убедительный аргумент в пользу достоверности, то результаты эксперимента определенно отдают предпочтение явлению A перед явлением B. Идеальной воспроизводимости нет ни в одном из двух случаев, но из глав 3 и 7 мы знаем, что не каждый эксперимент обязан быть успешным. Но все равно явление B воспроизводится только в половине экспериментов, поэтому стоило бы усомниться, действительно ли эффект существует.

Проблема в том, что явления A и B соответствуют реальным исследованиям. Явление A относится к так называемой прекогниции: способности человека получать информацию из будущего и использовать ее в настоящем. В статье, опубликованной в ведущем журнале в 2011 году, утверждалось, что в 9 из 10 экспериментов были получены значимые свидетельства прекогниции. Несмотря на опубликованные результаты, очень немногие ученые верят в существование преког-

ниции; в основном потому, что это подрывало бы весьма успешную общую теорию относительности. Таким образом, нам остается заключить, что высокого коэффициента воспроизводимости не всегда достаточно, чтобы люди поверили в достоверность эффекта.

Явление В относится к так называемому эффекту постороннего: люди, оказавшиеся свидетелями чрезвычайной ситуации и способные помочь, часто не пытаются помочь пострадавшим. Подобные эксперименты довольно трудно провести, потому что необходимы участники, выступающие в роли нуждающихся в помощи, и участники в роли свидетелей, не оказавших помощь. Поэтому исследования на эту тему имеют дело со сравнительно небольшими выборками. Как следствие, часто встречаются исследования эффекта постороннего, не дающие значимых результатов. И тем не менее почти все считают, что этот эффект реально существует. Стало быть, приходится заключить, что высокий коэффициент воспроизводимости не всегда необходим, чтобы заставить людей поверить в достоверность эффекта.

Похоже, мы оказались в парадоксальной ситуации. Ученые твердят о том, что воспроизводимость – золотой стандарт, позволяющий судить о достоверности эффекта, но при столкновении с *реальными* экспериментами воспроизводимость не является ни необходимым, ни достаточным условием достоверности. Возникает вопрос – зачем вообще ставить эксперименты?

Чтобы найти выход из этой странной ситуации, нужно лучше понимать, что такое статистика и воспроизводимость. В следующем разделе мы покажем, что эксперимент не всегда обязан допускать воспроизведение, особенно когда размеры эффекта и выборки малы. Успешное воспроизведение должно отражать оценочную вероятность успеха эксперимента. Беспокоиться следует, когда эксперимент воспроизводится слишком часто.

10.2. Тест избыточного успеха

Серия экспериментов должна завершаться успехом с частотой, близкой к вероятности успеха. Посмотрим, так ли это в экспериментах по изучению прекогниции.

Результаты каждого такого эксперимента анализировались с помощью одностороннего одновыборочного *t*-критерия. В табл. 10.1 приведены данные о размере выборки и стандартизованном размере эффекта (*g* Хеджеса) для каждого эксперимента. Воспользуемся техникой метаанализа из главы 9, чтобы вычислить объединенную оценку стандартизованного размера эффекта. Получится $g^* = 0.1855$.

Метаанализ здесь уместен, потому что автор исследования применял в том числе аналогичный анализ для обоснования своего теоретического заявления о существовании прекогниции. Объединенный размер эффекта, g^*, – лучшее, что мы можем предложить в качестве оценки размера эффекта, и его можно использовать для оценки мощности каждого отдельного эксперимента, как описано в главе 7. В последнем столбце табл. 10.1 приведена оценка мощности, основанная на таком метааналитическом размере эффекта. В полном соответствии с замечаниями по поводу мощности, сделанными в главе 7, мощность возрастает и убывает вместе с размером выборки. Ожидаемо в эксперименте 9 (n = 50) мощность наименьшая (0.36), а в эксперименте 7 (n = 200) – наибольшая (0.83). Примерно в половине экспериментов (где $n \approx 100$) мощность немного выше одной второй.

Таблица 10.1. Статистика для десяти экспериментов, которые должны были найти свидетельства в пользу прекогниции

	Размер выборки (n)	Размер эффекта (g)	Мощность
Эксп. 1	100	0.249	0.578
Эксп. 2	150	0.194	0.731
Эксп. 3	97	0.248	0.567
Эксп. 4	99	0.202	0.575
Эксп. 5	100	0.221	0.578
Эксп. 6a	150	0.146	0.731
Эксп. 6b	150	0.144	0.731
Эксп. 7	200	0.092	0.834
Эксп. 8	100	0.191	0.578
Эксп. 9	50	0.412	0.363

Предположим, что какой-то ученый решил воспроизвести эту серию из десяти экспериментов с такими же размерами выборок, как в оригинальной работе. Если мы согласны с тем, что объединенный размер эффекта – хороший показатель эффекта прекогниции, то ожидаемое число успешных исходов в серии из десяти экспериментов равно сумме их мощностей. Например, если мощность каждого эксперимента равна 1.0, то количество значимых результатов должно быть равно 10. Для экспериментов табл. 10.1 сумма мощностей

равна 6.27. Следовательно, эксперимент должен воспроизводиться 6.27 раз. Эта ожидаемая степень успеха заметно ниже, чем 9 из 10, указанная в оригинальной работе.

Какова вероятность получить 9 из 10 значимых результатов для этого эффекта? Мы можем проделать нечто подобное проверке гипотез и оценить вероятность получения 9 или более успешных результатов в 10 таких экспериментах. Необязательно, чтобы это были точно те же 9 экспериментов, что в статье, годятся любые 9 из 10. Чтобы вычислить вероятность успеха, найдем все 11 комбинаций экспериментов, в которых наблюдалось 9 или 10 успешных исходов. Для каждой комбинации вычислим вероятность именно такого результата, перемножив мощности всех успешных экспериментов и дополнения к мощностям всех неудачных. Затем сложим эти величины и получим 0.058. То есть если эффект действительно существует и близок к указанному в статье, то ученый, точно повторивший все десять оригинальных экспериментов, может рассчитывать всего лишь на 6%-ный шанс получить такой же успех, как в оригинальной статье. Если считать, что успешное воспроизведение должно быть основой доверия к достоверности экспериментальных результатов, то такой низкий коэффициент представляется серьезной проблемой.

Более того, низкая оценка коэффициента воспроизводимости поднимает вопрос о том, как автор оригинальной работы сумел получить такую высокую частоту успехов. С учетом того, что нам известно (из других исследований) об эффекте прекогниции, кажется очень странным, что 10 экспериментов оказались настолько успешными. Настолько странным, что можно заподозрить какие-то огрехи в методике проведения этой серии экспериментов. Возможно, мы никогда точно не узнаем, что случилось (этого может не знать и сам автор), но бремя доказательства лежит на исследователе, опубликовавшем результаты. Быть может, эффект прекогниции и существует, но это конкретное исследование не дает тому научного подтверждения.

Что, если применить такой же анализ к явлению B, в котором 10 из 19 экспериментов дали статистически значимые подтверждения эффекта постороннего? В этом случае объединенный стандартизованный размер эффекта равен –0.47, и отрицательное число свидетельствует о наличии эффекта. Объединенный размер эффекта можно использовать для оценки мощности каждого из 19 экспериментов. Мощность варьируется от 0.2 приблизительно до 1.0, потому что в нескольких экспериментах участвовало всего 24 человека,

а в одном было аж 2500 участников. Сумма мощностей всех экспериментов равна 10.77. Стало быть, следовало бы ожидать получения примерно 11 значимых результатов, а фактически в 10 экспериментах было получено 10 значимых результатов. Таким образом, серия экспериментов по исследованию эффекта постороннего представляется заслуживающей доверия, потому что частота успеха соответствует оценочной величине эффекта и размерам выборок. Оценочная вероятность наблюдать 10 или более значимых результатов в таком исследовании равна 0.76.

10.3. Избыточный успех как следствие статистического смещения публикации

В предыдущем разделе был описан тест избыточного успеха (ТИУ), который проверяет, согласуется ли заявленная автором частота успехов в серии экспериментов с оценочной величиной успеха и размерами экспериментальных выборок. Если расхождение велико, то ТИУ позволяет предположить какую-то проблему в постановке экспериментов, в анализе результатов или в теоретических утверждениях, основанных на данных или их анализе. В этом и следующем разделах мы на модельных экспериментах покажем, как коэффициент воспроизводимости может оказаться слишком большим. В этом разделе рассматривается влияние статистического смещения публикации: избирательной публикации значимых открытий и опускания незначимых.

В табл. 10.2 приведена статистика 20 модельных экспериментов, результаты каждого из которых анализировались с помощью двустороннего *t*-критерия. В каждом эксперименте была модельная контрольная группа, для которой никакого эффекта не было. Примеры для этой группы выбирались из нормального распределения с нулевым средним и единичным стандартным отклонением. Примеры для экспериментальных групп выбирались из нормального распределения со средним 0.3 и стандартным отклонением 1. Следовательно, стандартизованный эффект генеральной совокупности $\delta = 0.3$. Размеры выборок для обеих групп были одинаковы, $n_1 = n_2$. Они выбирались случайным образом из равномерного распределения в диапазоне от 15 до 50.

Во втором столбце табл. 10.2 показаны *t*-значения для каждого модельного эксперимента. Значения, выделенные полужирным шрифтом, статистически значимы, потому что соответствующие

p-значения меньше порога 0.05. Значимых экспериментов пять. Что ТИУ говорит о частоте успехов 5 из 20? Модельные данные можно трактовать так же, как в исследованиях прекогниции и эффекта постороннего. Объединив размеры эффектов во всех 20 экспериментах, мы получим $g^* = 0.303$. Эта оценка очень близка к истинному значению 0.3, так что мы имеем наглядную демонстрацию того, что метаанализ хорошо работает, когда в него включены все эксперименты. Объединенный размер эффекта можно использовать для оценки мощности каждого эксперимента; результаты показаны в четвертом столбце табл. 10.2. Сумма этих значений равна 4.2, а вероятность получения в таких экспериментах пяти или более значимых результатов равна 0.42. Не существует общепринятого порога вероятности успеха, но многие ощущают неудовлетворенность, если вероятность меньше 0.1. Если вклад в анализ вносят как значимые, так и незначимые результаты, то частота успехов примерно согласуется с оценочными значениями мощности. Пока все хорошо.

Теперь предположим, что автор публикации склонен к такой форме статистического смещения, когда публикуются и, следовательно, доступны для последующего изучения только значимые эксперименты (*t*-значения, выделенные полужирным шрифтом в табл. 10.2). Объединив только размеры эффектов в пяти опубликованных экспериментах, мы получим $g^* = 0.607$, т. е. удвоенный размер эффекта в генеральной совокупности. Это имеет смысл, потому что в значимых экспериментах *t*-значения должны быть сравнительно велики. Так как размер эффекта – функция от *t*-значения, в этих экспериментах должны быть также необычно большие оценочные размеры эффектов. Поэтому смещение публикации может стать причиной существенно завышенной оценки размера эффекта. Используя завышенный размер эффекта для вычисления мощности каждого эксперимента, получаем значения, показанные в последнем столбце табл. 10.2. Они намного больше истинных значений мощности, потому что основаны на сильно завышенных оценках размера эффекта. Тем не менее сумма мощностей равна 3.13, т. е. следует ожидать, что из пяти опубликованных экспериментов примерно три будут значимы. На самом деле значимые результаты были получены во всех пяти экспериментах, а вероятность, что все пять дадут значимые результаты, равна произведению мощностей, т. е. 0.081. Многие сочтут, что эта вероятность слишком низкая (например, меньше 0.1), и потому усомнятся в достоверности опубликованных результатов.

Таблица 10.2. Статистика для 20 модельных экспериментов с целью исследования эффекта статистического смещения публикации

$n_1 = n_2$	t	Размер эффекта	Мощность на основе объединенного РЭ	Мощность на основе смещенного РЭ
29	0.888	0.230	0.206	
25	1.380	0.384	0.183	
26	1.240	0.339	0.189	
15	0.887	0.315	0.126	
42	0.716	0.155	0.279	
37	1.960	0.451	0.251	
49	–0.447	–0.090	0.318	
17	1.853	0.621	0.138	
36	**2.036**	0.475	0.245	0.718
22	1.775	0.526	0.166	
39	1.263	0.283	0.262	
19	**3.048**	0.968	0.149	0.444
18	**2.065**	0.673	0.143	0.424
26	–1.553	–0.424	0.189	
38	–0.177	–0.040	0.257	
42	**2.803**	0.606	0.279	0.784
21	1.923	0.582	0.160	
40	**2.415**	0.535	0.268	0.764
22	1.786	0.529	0.166	
35	–0.421	–0.100	0.240	

10.4. Избыточный успех как следствие необязательной остановки

В главе 4 отмечалось, что одним из требований к *t*-критерию является фиксация размеров выборок в обеих группах до начала эксперимента. На практике, однако, очень часто бывает, что размер выборки *не* фик-

сирован. Рассмотрим следующую ситуацию. Исследователь собирает данные из двух генеральных совокупностей, и в результате в каждой выборке оказывается $n_1 = n_2 = 10$ примеров. Исследователь вычисляет t-критерий и находит, что $p = 0.08$. Это p-значение выше порога статистической значимости 0.05, но все равно выглядит многообещающе. В таком случае исследователи часто решают взять еще десять примеров, чтобы в каждой выборке было $n_1 = n_2 = 20$ примеров. Предположим, что при вычислении t-критерия на большей выборке получается $p = 0.04$, это уже ниже порога статистической значимости. Выглядит прекрасно: взяли больше данных – получили более точный ответ. К несчастью, такая процедура может сильно увеличить частоту ошибок типа I. Одна из проблем заключается в том, что в процедуре участвует несколько проверок. И с каждой из них связана какая-то вероятность допустить ошибку типа I. Как показано в главе 5, если выполняется несколько проверок, то вероятность, что хотя бы в одной из них будет допущена ошибка типа I, выше, чем вероятность ошибки типа I в одной проверке.

Еще более серьезная проблема заключается в том, что сбор данных прекращается после достижения желаемого результата. По мере добавления наблюдений в первоначальный набор данных значимость может уступать место незначимости, и наоборот. Если решение о добавлении данных увязывается с получением значимого результата (т. е. данные перестают добавляться, как только $p < 0.05$), то процесс сбора данных оказывается смещен в сторону получения значимых результатов. Такая процедура называется «необязательной остановкой», она увеличивает частоту ошибок типа I. Недобросовестный исследователь, который начинает с $n_1 = n_2 = 10$ и добавляет по одному наблюдению в каждый набор данных, пока не получится значимый результат ($p < 0.05$) или до максимума $n_1 = n_2 = 50$, будет вознагражден частотой ошибок типа I, превышающей 20 %.

Важно понимать, что проблема здесь не в *добавлении* данных, а в *остановке* сбора данных, потому что частота ошибок типа I относится к *полной* процедуре. Следовательно, необязательная остановка является проблемой, даже если первый набор данных дает значимый результат, но исследователь *мог бы* добавить больше примеров в незначимый набор данных. Важно, что если у исследователя нет конкретного плана сбора данных, то невозможно вычислить частоту ошибок типа I. Именно поэтому стандартный подход к проверке гипотез предполагает, что размер выборки фиксирован.

ТИУ чувствителен к серии экспериментов, в которой исследователь применяет такой некорректный подход, и, чтобы лучше прочувствовать, что происходит, очень полезно взглянуть на модельные эксперименты. В табл. 10.3 приведена статистика для 20 модельных экс-

периментов, результаты которых анализировались с помощью двухвыборочного *t*-критерия. Размеры выборок в контрольной и экспериментальной группах, n_1 и n_2, были одинаковы. Примеры выбирались из нормального распределения с нулевым средним и единичным стандартным отклонением. Поэтому размер эффекта в генеральной совокупности $\delta = 0$, т. е. эффекта в действительности не существует.

Таблица 10.3. Статистика по 20 модельным экспериментам для исследования эффектов необязательной остановки

$n_1 = n_2$	*t*	Размер эффекта	Мощность на основе объединенного РЭ	Мощность на основе значимых экспериментов
19	**2.393**	0.760	0.053	0.227
100	0.774	0.109	0.066	
100	1.008	0.142	0.066	
63	**2.088**	0.370	0.060	0.611
100	0.587	0.083	0.066	
100	−1.381	−0.195	0.066	
100	−0.481	−0.068	0.066	
100	0.359	0.051	0.066	
100	−1.777	−0.250	0.066	
100	−0.563	−0.079	0.066	
100	1.013	0.143	0.066	
100	−0.012	−0.002	0.066	
46	**2.084**	0.431	0.057	0.480
100	0.973	0.137	0.066	
100	−0.954	−0.134	0.066	
100	−0.136	−0.019	0.066	
78	**2.052**	0.327	0.062	0.704
100	−0.289	−0.041	0.066	
100	1.579	0.222	0.066	
100	0.194	0.027	0.066	

Полужирным шрифтом выделены статистически значимые результаты ($p < 0.05$)

Для моделирования необязательной остановки в каждую выборку первоначально было включено $n_1 = n_2 = 15$ примеров. К данным применялся t-критерий и при обнаружении незначимого результата в каждую группу добавлялся еще один пример, затем t-критерий вычислялся снова. Этот процесс повторялся, пока размеры выборок не достигали $n_1 = n_2 = 100$, после чего выдавался окончательный результат.

Поскольку размер эффекта в генеральной совокупности равен нулю, следует ожидать, что в среднем будет один значимый результат на 20 модельных экспериментов (см. главу 5). Четыре выделенных полужирным шрифтом t-значения в табл. 10.3 соответствуют статистической значимости; эта частота (20 %) гораздо выше ожидаемых 5 %. Простое вычисление с использованием биномиального распределения показывает, что вероятность получить четыре или более значимых результатов в серии из 20 экспериментов равна 0.016, если каждый эксперимент дает значимый результат с вероятностью 5 %. Во всех незначимых экспериментах в табл. 10.3 размер выборки равен 100 (максимально возможный), потому что такова природа необязательной остановки.

Вычисление объединенного размера эффекта для всех 20 экспериментов дает $g^* = 0.052$, очень близко к размеру эффекта в генеральной совокупности, равному нулю. В отличие от смещенной публикации необязательная остановка не приводит к смещенным оценкам размера эффекта. И если использовать этот оценочный размер эффекта для вычисления мощности каждого эксперимента, то получатся значения в диапазоне от 0.053 до 0.066; все они лишь немного превышают порог значимости 0.05, потому что оценочный размер эффекта лишь чуть больше нуля. И тем не менее результаты кажутся слишком хорошими, чтобы быть правдой. Сумма мощностей во всех 20 экспериментах равна всего 1.28, поэтому мы ожидаем найти примерно один значимый эксперимент из 20. Вычисленная по этим значениям мощности вероятность того, что среди таких экспериментов будет четыре и более значимых, равна 0.036. Этот результат (правильно) указывает на наличие какой-то проблемы в серии экспериментов: частота успехов выше, чем должна быть.

В последнем столбце табл. 10.3 приведены значения мощности, основанные только на значимых экспериментах. Предполагается, что информация о незначимых не опубликована (статистическое смещение публикации). В этой ситуации анализ ТИУ вынужден работать только с четырьмя значимыми экспериментами. Оценка объединенного размера эффекта $g^* = 0.4$ – намного выше истинного значения 0. В результате этого завышения оценки значения мощно-

сти четырех значимых экспериментов также сильно переоценены. И тем не менее сложение этих мощностей показывает, что среди четырех таких экспериментов должно быть примерно два значимых. Вероятность, что все четыре эксперимента дадут значимые результаты, равна произведению мощностей, т. е. 0.047. И снова серия экспериментов (правильно) кажется сомнительной, потому что частота успехов не согласуется с оценкой размера эффекта и размерами экспериментальных выборок.

10.5. Избыточный успех и теоретические утверждения

Тест избыточного успеха может выявить ситуации, когда опубликованная частота успехов не согласуется с экспериментальными размерами эффекта и выборки. В этом анализе важный момент – определение «успеха», которое всегда соотносится с каким-то теоретическим утверждением. Предположим, к примеру, что исследователь поставил десять независимых экспериментов, в каждом из которых изучается отдельный вопрос (например, эффект Струпа, эксперимент по запоминанию, различия в альфа-ритме на ЭЭГ, эпигенетическая передача поведения, приобретенного в результате обучения, прекогниция и другие вопросы). Предположим, что в первых четырех экспериментах получен значимый результат, а в остальных шести – нет. Предположим также, что исследователь опубликовал только четыре успешных результата, опустив шесть нулевых. Анализ ТИУ по четырем опубликованным исследованиям (правильно) указывает на признаки статистического смещения публикации, но это наблюдение не имеет особого смысла. Все четыре эксперимента не связаны между собой и не подразумевают какого-то общего теоретического утверждения. А раз так, то мы можем лишь заключить, что были какие-то неудачные эксперименты, о которых не сообщается, но само их существование ничего не говорит о достоверности опубликованных свойств эффекта Струпа или эффективности запоминания.

С другой стороны, если тот же исследователь использовал результаты тех же самых четырех значимых экспериментов, чтобы выдвинуть какое-то теоретическое утверждение (например, объединенную теорию эффекта Струпа, запоминания, альфа-ритмов и эпигенетической передачи), то статистическое смещение публикации потенциально ставит это утверждение под сомнение. Если анализ ТИУ показывает, что серия четырех публикаций может быть смеще-

на, то ученые должны скептически отнестись к теоретическим заявлениям, в основе которых лежат эти эксперименты.

Часто исследователи непреднамеренно формулируют теоретические выводы, слишком хорошие, чтобы быть правдой, поскольку кладут в основу теории значимость и незначимость выполненных экспериментов. В таком случае теория становится всего лишь кратким и грубым изложением того, что было измерено в эксперименте. Такая теория почти наверняка объясняет какой-то шум в экспериментальных результатах и, скорее всего, не будет подтверждена новой серией экспериментов.

В полном соответствии со следствием 3а в главе 3 выводы анализа ТИУ не доказывают отсутствия эффекта в серии экспериментов, а говорят лишь, что в этой серии нет убедительной научной аргументации.

Что следует запомнить

1. Если каждый из многих похожих экспериментов с малыми размерами эффекта и выборки дает значимый результат, то данные слишком хороши, чтобы быть правдой.
2. Количество значимых результатов в экспериментах должно быть пропорционально их мощности.
3. Статистическое смещение публикации и необязательная остановка могут сильно увеличить частоту ошибок типа I.

Литература

1. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349. *https://doi.org/10.1126/science.aac4716*.

Глава 11

Величина избыточного успеха

Что вы узнаете из этой главы

В главе 10 мы познакомились с тестом избыточного успеха (ТИУ), который обнаруживает некоторые виды смещения в статистическом анализе нескольких экспериментов. Хотя мы и нашли серии экспериментов, для которых ТИУ указывал на наличие проблем, можно было бы допустить, что подавляющее большинство научных исследований этим не грешит. Но, как показано ниже, это, к сожалению, не так.

11.1. При определении смещения возможны трудности

Основная идея ТИУ довольно проста. Важное положение статистики заключается в том, что неудачи обязательно должны быть. Даже если эффект реально существует, иногда должны встречаться выборки, в которых он не проявляется. Сообщать только об успешных результатах чревато проблемами, потому что опубликованный размер эффекта может оказаться завышенным, и это указывает на существование эффекта, которого на самом деле нет. Слишком высокий коэффициент воспроизводимости – признак наличия какой-то проблемы в процессе сбора данных, анализа, выдвижения теории или публикации.

Чтобы продемонстрировать последствия такого рода интерпретации, рассмотрим три серии модельных результатов, показанные в табл. 11.1. Каждый смоделированный набор данных анализировался с помощью двустороннего *t*-критерия. В основе каждой серии из пяти исследований лежали по-разному смоделированные эксперименты. В одной серии (корректная проверка) размер эффекта в генеральной совокупности был равен 0.8. Размеры выборок, $n_1 = n_2$, случайно выбирались из диапазона от 10 до 30. Все пять экспериментов дали значимые результаты, которые были полностью опубликованы.

Таблица 11.1. Сводная статистика для трех серий из пяти модельных экспериментов

Серия A				Серия B				Серия C			
$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g	$n_1 = n_2$	t	p	g
10	2.48	0.03	1.06	21	2.67	0.01	0.81	16	2.10	0.04	0.72
28	2.10	0.04	0.55	27	4.72	<0.01	1.26	19	2.19	0.04	0.70
10	3.12	0.01	1.34	22	3.66	<0.01	1.08	25	2.22	0.03	0.62
15	2.25	0.04	0.80	26	2.74	0.01	0.75	14	2.24	0.04	0.82
12	2.34	0.03	0.92	24	2.06	0.05	0.58	23	2.49	0.02	0.72

При генерировании одной серии была использована необязательная остановка, при генерации другой – смещение публикации, а в третьей серии все хорошо. В какой серии нет подвохов?

Еще одна серия исследований в табл. 11.1 основана на модельных экспериментах, в которых размер эффекта в генеральной совокупности равен 0 (эффект отсутствует). Выборка генерировалась с применением необязательной остановки: вначале $n_1 = n_2 = 10$, а затем размер увеличивался с шагом 1 до достижения максимума 30. Всего было смоделировано 20 экспериментов, пять из которых дали значимый результат. Только эти пять значимых экспериментов и были опубликованы.

И третья серия исследований в табл. 11.1 основана на модельных экспериментах, в которых истинный размер эффекта равен 0.1. Размер выборки случайно выбирался из диапазона от 10 до 30. Всего было смоделировано 100 экспериментов, из которых пять дали значимый результат. Только они и были опубликованы, а остальные 95 были опущены как незначимые.

Читателю предлагается определить, как генерировалась каждая серия экспериментов в табл. 11.1. Чтобы не было никакой неясности, повторим, что в одной серии экспериментов размер эффекта в генеральной совокупности очень велик, и исследовался он на всех пяти опубликованных экспериментах. Это корректная серия. Во второй серии экспериментов эффекта не было вовсе, но применялась необязательная остановка, а публикация была статистически смещена, поскольку публиковались только значимые результаты. Это некорректная серия. В третьей серии размер эффекта очень мал, экспериментов проведено много, но опубликовано только пять значимых результатов. Эта серия также некорректна. Так какая же серия в табл. 11.1 корректна? Мы предлагаем читателю изучить статистику и принять решение, прежде чем читать дальше.

Нашли корректную серию? Если задание показалось вам трудным или вы не уверены в правильности ответа, утешьтесь – вы не одиноки. Даже ученые с солидным опытом часто испытывают сложности при оценке этих статистических данных. Давайте остановимся на следствии из этого наблюдения. Если возможны смещение публикации и необязательная остановка, то ученые не знают, как отличить серии экспериментов, в которых эффекта не было вообще, от серий, в которых эффект очень велик.

Метаанализ с объединением опубликованных размеров эффектов дает: для серии A $g^* = 0.82$, для серии B $g^* = 0.89$, а для серии C $g^* = 0.70$. Но и эта информация многим ученым не помогает идентифицировать корректную серию экспериментов.

Здесь оказывается полезен тест избыточного успеха. Детали вычислений оставим в качестве упражнения для читателя, но если вычислить мощность для каждой серии и перемножить мощности, то получится, что вероятность, с которой все пять экспериментов дают значимые результаты, равна: для серии A $p = 0.042$, для серии B $p = 0.45$, для серии C $p = 0.052$. И действительно, корректной является серия B.

ТИУ – это формальный анализ, но существуют эвристические правила, которые позволяют быстро оценить корректность серии экспериментов. В частности, можно посмотреть на связь между размером выборки и размером эффекта. В корректной серии эти величины не связаны (больший размер выборки ведет к более точным оценкам размера эффекта, но на сам размер эффекта не влияет). Для серии B в табл. 11.1 корреляция между размером выборки и размером эффекта $r = 0.25$, что отражает случайность размера эффекта в разных экспериментах. Напротив, для серий A и C $r = -0.86$ и $r = -0.83$ соответственно. Эту связь легко объяснить в случае, когда используется необязательная остановка: выборка может быть велика, только если размер эффекта мал (если бы оценочный размер эффекта был велик, то уже малая выборка дала бы значимый результат). Подобные связи имеются и в других сериях экспериментов; например, в исследованиях, ставящих целью найти подтверждение прекогниции (табл. 10.1 в главе 10), корреляция между размером выборки и размером эффекта $r = -0.89$.

Еще один признак проблематичного набора данных – когда многие p-значения близки к порогу статистической значимости, но всегда остаются ниже него. В экспериментах с необязательной остановкой очень часто получаются статистики с p-значением чуть ниже порогового. Напротив, в корректных экспериментах с реальным эффектом и подходящими размерами выборок обычно получаются очень малые

p-значения, тогда как значения, близкие к порогу, должны встречаться редко. Легко видеть, что в серии B в табл. 11.1 почти все *p*-значения очень малы, а в других сериях имеется много *p*-значений от 0.02 до 0.05. Такое распределение *p*-значений должно насторожить: в этой серии экспериментов есть что-то странное.

11.2. Насколько широко распространены эти проблемы?

Пока что мы установили, что результаты некоторых серий экспериментов выглядят слишком хорошо, чтобы быть правдой, и что такие факты должны поставить под сомнение нашу уверенность в достоверности выводов автора. Но само существование проблематичных экспериментов еще не говорит, что такого рода проблемы широко распространены; вполне возможно, что они встречаются редко. Несмотря на сомнения по поводу некоторых исследований, не хотелось бы без нужды подозревать целую область науки.

Чтобы исследовать, насколько распространены такие проблемы, можно, например, систематически проанализировать конкретный набор исследований. Science – один из ведущих научных журналов, у него свыше 100 000 подписчиков, и это главный плацдарм, который должен завоевать любой молодой ученый, рассчитывающий на получение должности ассистента профессора или на утверждение пожизненного контракта. Хочется надеяться, что уж такой-то журнал публикует только самые лучшие работы в любой области, особенно если учесть, что процент одобренных к публикации работ очень низок – около 7 %. Онлайновая система поиска по журналу нашла 133 статьи по психологии или образованию, опубликованных в период между 2005 и 2012 годом. Мы применили анализ ТИУ к каждой из 18 статей, в которых было описано четыре и более эксперимента и содержалось достаточно информации для оценивания вероятности успеха.

В табл. 11.2 приведены оценки вероятностей успеха для этих 18 сследований. Удивительно, но в 15 из 18 (83 %) статей в Science результаты оказались слишком хорошими, чтобы быть правдой (т. е. вероятность успеха меньше 0.1). Вполне возможно, что читатель узнает кое-какие статьи по кратким названиям, потому что многие результаты были описаны в популярной литературе, а некоторые стали основой для принятия стратегических решений в области образования, благотворительности и соблюдения диеты.

Таблица 11.2. Результаты анализа ТИУ для статей в журнале Science

Год	Краткое название	Вероятность успеха
2006	Deliberation-Without-Attention Effect	0.051
2006	Psychological Consequences of Money	0.002
2006	Washing Away Your Sins	0.095
2007	Perception of Goal-Directed Action in Primates	0.031
2008	Lacking Control Increases Illusory Pattern Perception	0.008
2009	Effect of Color on Cognitive Performance	0.002
2009	Monkeys Display Affiliation Toward Imitators	0.037
2009	Race Bias via Televised Nonverbal Behavior	0.027
2010	Incidental Haptic Sensations Influence Decisions	0.017
2010	Optimally Interacting Minds	0.332
2010	Susceptibility to Others' Beliefs in Infants and Adults	0.021
2010	Imagined Consumption Reduces Actual Consumption	0.012
2011	Promoting the Middle East Peace Process	0.210
2011	Writing About Worries Boosts Exam Performance	0.059
2011	Disordered Contexts Promote Stereotyping	0.075
2012	Analytic Thinking Promotes Religious Disbelief	0.051
2012	Stop Signals Provide Inhibition in Honeybee Swarms	0.957
2012	Some Consequences of Having Too Little	0.091

Одно исследование в табл. 11.2 («Disordered Contexts Promote Stereotyping»[1]) заслуживает специального обсуждения. Поставленный на первое место автор, Дидерик Стапель, голландский специалист по социальной психологии, был уличен в фальсификации данных. И действительно, данные, на которые опирается эта статья, были собраны не в реальном эксперименте, а сгенерированы

[1] «Беспорядок в окружающей среде ведет к стереотипным суждениям». – *Прим. перев.*

с помощью электронной таблицы ведущим автором (второй автор не знал о фальсификации). Возможно, вы думаете, что фальсификатор должен был позаботиться о том, чтобы данные выглядели правдоподобными, но опубликованные (поддельные!) результаты казались слишком хорошими, чтобы быть правдой. Очень может статься, что Стапель сгенерировал сфальсифицированные данные, похожие на реальные данные из опубликованных экспериментов; к сожалению, опубликованные (предположительно реальные) данные часто выглядят слишком хорошо, чтобы быть правдой.

Закономерности, обнаруженные для статей в Science, не уникальны. ТИУ-анализ статистики для статей в журнале Psychological Science выявил схожую пропорцию избыточного успеха (36 из 44 статей, 82 %, казались слишком хорошими, чтобы быть правдой). Похоже, проблема свойственна не только психологии, в некоторых статьях по эпигенетике и нейронаукам наблюдаются аналогичные проблемы.

Из табл. 11.2 и других подобных исследований вытекает общий вывод: ведущие ученые, редакторы и рецензенты не понимают, как выглядят добротные научные данные, когда исследование включает несколько экспериментов и проверок. В лучшем случае вполне вероятно, что многие считающиеся эталонными экспериментальные работы по психологии и другим дисциплинам, опирающимся на статистику, с помощью подобных экспериментов и анализа их результатов доказывают невоспроизводимое.

11.3. Что происходит?

Сейчас уместно будет сделать шаг назад и поговорить о том, как наука оказалась в таком положении. Конечно, ученые испытывает мощное давление, они обязаны публиковать результаты успешных экспериментов (а как иначе получить место на кафедре или грант?), но большинство (по крайней мере, многие) искренне любят свою область науки и верят в то, что публикуют достоверные и важные открытия, которые могут принести пользу обществу. Стало быть, приходится заключить, что многие ученые не понимают роль статистического анализа в интерпретации эмпирических данных. Следующее далее обсуждение по необходимости умозрительное, но нам кажется, что имеет смысл обсудить некоторые распространенные недоразумения.

11.3.1. Непонимание воспроизводимости

В главе 10 мы отмечали, что успешное воспроизведение часто считается «золотым стандартом» в научной работе. Но вот чего многие

ученые недопонимают, так это то, что в правильно поставленной серии экспериментов частота успешного воспроизведения должна соответствовать экспериментальной мощности (более общо – вероятностям успеха). Упор на успех воспроизведения застит ученым глаза и мешает заметить, что опубликованные эксперименты с умеренной или низкой мощностью тем не менее почти всегда оказывались успешными. Но просто в силу случайности выборки эксперименты с низкой мощностью не должны всякий раз быть успешными. Перечислим несколько причин, по которым исследования с низкой мощностью так часто дают значимые результаты, хотя и не должны бы.

11.3.2. Статистическое смещение публикации

Ученый может опубликовать результаты эксперимента, подкрепляющие некоторое теоретическое положение, и не публиковать опровергающие его результаты. Если бы каждый эксперимент давал четкий ответ на научный вопрос, такое поведение следовало бы назвать фальсификацией. Однако зачастую трудно понять, «сработал» ли эксперимент. Учитывая сложность многих экспериментов (например, клеточную культуру нужно должным образом вырастить, прежде чем можно будет говорить о тормозящем эффекте исследуемого химического вещества), есть масса причин неудачи эксперимента. Ученый может счесть, что в эксперименте, не показавшем желаемый результат, была допущена какая-то методологическая ошибка, подспудно не желая признать, что он дал отрицательный ответ на поставленный вопрос. За некоторыми учеными числятся многочисленные исследования, ошибочно названные «пилотными», хотя на самом деле их следовало бы трактовать как отрицательные ответы.

11.3.3. Необязательная остановка

Эмпирические науки постоянно нуждаются в дополнительных данных. Этот подход имеет свою ценность, но часто вступает в конфликт с характеристиками проверки гипотез. Например, в главе 10 мы отмечали, что необязательная остановка увеличивает частоту ошибок типа I. Эту проблему очень трудно решить в рамках проверки гипотез. Предположим, к примеру, что ученый фиксирует близкий к порогу (p = 0.07) результат в эксперименте 1 и решает провести новый эксперимент 2 для проверки эффекта. Может показаться, что ученый скрупулезно делает свою работу, но на самом деле это необязательно. Допустим, что эксперимент 1 показал значимый эффект (p = 0.03), стал бы тогда ученый ставить эксперимент 2 для проверки? Если нет, значит, по существу, действия ученого не что

иное, как необязательная остановка на уровне нескольких экспериментов, а частота ошибок типа I в любом конкретном эксперименте (или на множестве экспериментов) неизвестна.

В действительности проблема необязательной остановки связана не с фактическими действиями ученого (например, исследование с запланированным размером выборки дает $p = 0.02$), а тем, что он сделал бы, если бы результат оказался иным (например, если исследование с запланированным размером выборки дает $p = 0.1$, добавил бы он еще 20 субъектов?). Точнее, если вы не знаете, что стали бы делать при любом возможном развитии событий, то не можете знать, какова частота ошибок типа 1 в вашем анализе.

11.3.4. Выдвижение гипотез после того, как результаты стали известны

В некоторых научных исследованиях бывает так, что ученый собирает данные из многих экспериментов, а затем пытается сочинить складную историю, которая связала бы разрозненные результаты воедино. Этот подход выглядит как добросовестная научная практика, потому что ученый не отходит от данных, но на самом деле порождает теории, которые *слишком* близки к данным и в результате объясняют не только сигнал, но и шум. Почти всегда можно придумать апостериорную теорию, объясняющую, почему сигнал появился или исчез. И наоборот, данные, не укладывающиеся в теорию, можно назвать несущественными и обоснованно (по мнению ученого) исключить из серии экспериментов.

Такое апостериорное рассуждение применимо к измерениям в рамках как одного, так и нескольких экспериментов. Ученый может произвести несколько измерений, найти то, которое, как ему кажется, дает согласованные результаты в нескольких экспериментах, и заключить, что это измерение наилучшее. И такая практика может показаться добросовестной (и является таковой, если применять ее правильно), но часто ведет к выбору одного измерения на основе случайной вариативности выборки. Другие измерения могут быть ничуть не хуже (или даже лучше), но, увы, не показали эффекта (а, может быть, совершенно правильно показали, что эффекта и не было).

11.3.5. Гибкость анализа

Современные программы позволяют выполнять широкий спектр анализов в поисках статистической значимости. Данные не показывают значимого результата? А давайте возьмем от них логарифмы или обратные значения и попробуем еще раз. Все равно нет значи-

мости? Давайте-ка исключим выбросы, отличающиеся от среднего на три или на 2,5 стандартных отклонения. Исключим эффекты пола и потолка (глава 2) или уберем данные, полученные для участников, не удовлетворяющих еще какому-то критерию. Если в эксперименте измеряется несколько показателей, их можно скомбинировать самыми разными способами (усреднить, взять максимум, перемножить, выполнить анализ главных компонент). Хотя исследование данных – замечательное упражнение для ученого на этапе предварительной проработки, оно увеличивает частоту ошибок типа I в регулярном эксперименте. Поэтому, если вы пробовали различные способы анализа данных и обнаружили какой-то значимый результат, то необходимо повторить эксперимент на независимой выборке и применить к полученным данным тот же вид анализа.

Стандартные виды анализа, похоже, поощряют такую гибкость. Напомним, что в разделе 6.7 было показано, что ANOVA типа 2×2 с вероятностью 14 % дает по меньшей мере один значимый результат (основной эффект или взаимодействие) для нулевых данных. Этому свойству можно противопоставить хорошее понимание того, какие именно тесты пригодны в вашем исследовании. А затем (с полным основанием) игнорировать результаты других тестов, сообщенные ANOVA.

11.3.6. Непонимание того, что такое предсказание

Научные аргументы кажутся особенно убедительными, когда теория предсказывает неизвестный ранее результат, который затем проверяется на экспериментальных данных. На самом деле во многих статьях в Science, проанализированных в табл. 11.2, встречаются фразы типа «как и предсказано теорией, обнаружилось значимое различие». Такие утверждения звучат странно на двух уровнях. Во-первых, даже если эффект существует, проверка гипотез не будет давать значимый результат каждый раз. В лучшем случае теория может лишь предсказать вероятность значимого результата для выборки данного размера. Во-вторых, чтобы теория предсказывала вероятность успеха (обычно это мощность), она должна указывать размер эффекта при данном плане эксперимента и размерах выборок. Ни в одной статье из перечисленных в табл. 11.2, не было обсуждения теоретически предсказанного размера эффекта или мощности.

Это означает, что фраза «как и предсказано теорией, обнаружилось значимое различие» в этих статьях бессодержательна. Теория, может, и существует, но это не та теория, которая способна предсказать вероятность получения значимого результата в эксперименте. (Надо полагать, что если бы теория была на это способна, то ее авто-

ры не упустили бы возможность обсудить детали.) Таким образом, фактически никакого предсказания и нет. Складывается странная ситуация, когда теоретические предсказания, которые предсказаниями вовсе не являются, неизменно оказываются правильными. И свидетельствуют об успехе там, где его достижение должно быть принципиально невозможной задачей.

11.3.7. Небрежность и избирательная двойная проверка

В таком сложном виде деятельности, как наука, ошибки неизбежны. Ошибки ввода данных, ошибки вычислений, ошибки при копировании и вставке – все это может стать причиной неверной интерпретации. Хотя ученых учат все проверять и перепроверять, ошибки такого типа встречаются в опубликованных работах очень часто. Компьютерная программа STATCHECK умеет анализировать опубликованные статьи и проверять, осмыслена ли приведенная в них статистика. Например, если в статье говорится, что $t(29) = 2.2$, $p = 0.01$, то ошибка определенно присутствует, потому что значениям $t = 2.2$ и $df = 29$ соответствует $p = 0.036$. В ходе анализа тысяч статей по психологии STATCHECK обнаружила, что примерно в половине имеется по меньшей мере одна ошибка такого рода. А почти в 10 % статей имеется по меньшей мере одна ошибка, из-за которой результат перестал быть значимым.

Подобные ошибки в опубликованных работах могут указывать на общую небрежность при проведении исследований. Такая невнимательность может означать, что даже добросовестные исследователи публикуют не заслуживающие доверия результаты. Хуже того, степень небрежности при обработке данных может зависеть от того, согласуются ли опубликованные результаты с надеждами или ожиданиями автора. Например, если из-за какой-то ошибки при вводе данных t-критерий показал, что $t(45) = 1.8$, $p = 0.08$, то исследователь, возможно, перепроверит процедуру ввода данных, найдет ошибку и, повторив анализ, получит $t(45) = 2.3$, $p = 0.03$. С другой стороны, если ошибка привела к значимому результату, например $t(52) = 2.4$, $p = 0.02$, то исследователь может и пренебречь перепроверкой процедуры ввода данных, хотя при этом, возможно, обнаружил бы ошибку и получил результат $t(52) = 1.7$, $p = 0.1$.

Поскольку в научных исследованиях масса мест, где может возникнуть ошибка, ученый вполне может непреднамеренно исказить свои результаты, избирательно перепроверяя нежелательные исходы и доверяя желательным.

Что следует запомнить

1. Во многих областях науки, включая медицину, биологию, психологию и др., воспроизводимость слишком высока.
2. Создается впечатление, что многие ученые применяют методы, от которых лучше бы держаться подальше: необязательная остановка, статистическое смещение публикации, выдвижение гипотез после того, как результаты стали известны, чрезмерная гибкость анализа и др.

Глава 12

Предлагаемые улучшения и нерешенные проблемы

Что вы узнаете из этой главы

Тест избыточного успеха высветил проблемы, имеющиеся в современной научной практике в тех областях, где используется проверка гипотез. Кроме того, часто встречаются ошибки в опубликованных статистических данных (например, *p*-значение не соответствует *t*-значению), результаты, доложенные на конференции, изменяются в журнальной публикации, а исследователи прибегают к статистическому смещению публикации и другим порицаемым практикам. Эти проблемы побудили многих ученых предложить контрмеры, призванные улучшить научную работу. В этой главе мы дадим критический обзор некоторых предложений такого рода. Хотя во многих предложениях можно найти плюсы, зачастую имеются и минусы, и, похоже, ни одно предложение не затрагивает фундаментальных проблем. У нас нет какого-то одного предложения, решающего все проблемы разом, но мы определили, что, на наш взгляд, должно стать важнейшими долгосрочными целями науки, и предлагаем практические меры, которые должны поспособствовать достижению этих целей.

12.1. Любой ли эксперимент следует публиковать?

Некоторые считают, что ученый обязан публиковать любой экспериментальный результат независимо от статистической значимости. Действительно, незначимые эксперименты содержат информацию об эффектах (нулевую или нет), которую можно использовать (с помощью метааналитических методов), если данные опубликованы. Вообще, публикация всех данных дает читателям возможность делать правильные выводы об эффектах и избегать переоценки размеров эффектов, вызванной статистическим смещением публикации.

Однако публикация всех экспериментальных результатов сопряжена с некоторыми трудностями. Во-первых, может ли экспериментатор надежно отличить эксперимент, потерпевший неудачу по методологическим причинам (например, из-за поломки оборудования), от эксперимента, не принесшего успеха из-за случайности выборки? Если эти разные типы неудач неразличимы, то литература окажется замусоренной экспериментальными результатами, недостоверными по самым разным причинам (конечно, не исключено, что это происходит уже сейчас, но публикация всего на свете может усугубить проблему).

Кроме того, непонятно, как ученые, публикующие все результаты, должны интерпретировать свои открытия. Может ли вообще присутствовать раздел «Выводы» в статье, которая просто увеличивает объем данных по уже существующей теме? А как решить, какие результаты включать при выполнении метаанализа? Все эти проблемы существуют уже сейчас, но публикация всех данных не снимет их, а только обострит.

Наконец, когда можно заключить, что по некоторому вопросу уже собрано достаточно данных? Если метаанализ дает $p = 0.08$, то следует ли ставить дополнительные эксперименты, пока не окажется, что $p < 0.05$? Такой подход был бы проблематичным из-за необязательной остановки только на уровне экспериментов, добавляемых в метаанализ, а не на уровне индивидуальных субъектов, добавляемых в один эксперимент. Настанет ли когда-то момент, когда сообщество ученых может решить, что эффект существует (см. главу 3, следствие 1а)? А что, если дополнительные данные способны изменить решение?

12.2. Предварительное объявление

Некоторые журналы стали поощрять и продвигать контрольные исследования и часто требуют, чтобы исследователи предварительно объявляли о своем эксперименте, методах анализа и плане сбора данных. Есть ученые, которые считают, что предварительное объявление – единственный жизнеспособный способ вывести психологию (и другие науки, в которых используется статистика) из состояния, которое им представляется кризисом.

Идея предварительного объявления заключается в том, что перед реальным проведением эксперимента ученые описывает весь его план в таком месте, которое лишило бы его возможности впоследствии изменить исходный план (например, в рамках проекта Open Science Framework или на сайте AsPredicted.org). В плане должны

быть описаны стимулы, задания, экспериментальные методы, количество выборок и способ их формирования, вопросы обследуемым и план анализа данных. После документирования всех этих деталей эксперимент проводится, и все отклонения от заранее объявленного плана отмечаются (быть может, с обоснованием). Сторонники предварительного объявления говорят, что этот подход помешает исследователям выдвигать теоретические идеи или изменять методы анализа данных после знакомства с данными. При наличии предварительного объявления станет очевидно, что исследователь слишком рано прекратил сбор данных или добавил наблюдения (быть может, вследствие необязательной остановки) или что различные измерения были скомбинированы способом, отличающимся от запланированного. Если все предварительные объявления хранятся в месте, открытом всем желающим, то это может также уменьшить число публикаций со статистическим смещением, потому что имеется публичная запись о том, какой эксперимент собирался поставить исследователь. Из тех же соображений журналы могли бы согласиться публиковать предварительно объявленные эксперименты еще до сбора данных.

Все эти аргументы кажутся вескими прагматическими причинами перехода к практике предварительного объявления. Но стоит задуматься, как возникают вопросы, что случится, если исследователь будет строго придерживаться предварительно объявленного плана. Дает ли приверженность этой стратегии дополнительную уверенность в результатах или теоретических выводах? Повышается ли наше доверие к процессу, в результате которого на свет появился предварительно объявленный план эксперимента? Рассмотрение двух крайних случаев показывает, что нет.

Крайний случай 1. Предположим, что исследователь выдвигает гипотезу, подбрасывая монету. Например, лекарственное средство может увеличить или уменьшить объем кратковременной памяти. Выпадает орел, и исследователь предварительно объявляет гипотезу о том, что под воздействием лекарства объем памяти увеличивается. Затем он проводит эксперимент, который подтверждает предсказанный эффект. Неважно, различаются генеральные совокупности или нет, очевидно, что такой исход эксперимента ни в коей мере не является обоснованием процесса выдвижения гипотезы (подбрасывание монеты). Чтобы эксперимент подтверждал предсказание (а не просто гипотезу), должно быть какое-то обоснование теории или процесса

его формирования. Предварительное объявление не дает и не может дать такого обоснования, поэтому объявлять о ничем не обоснованном плане эксперимента как-то глупо.

Крайний случай 2. Предположим, что исследователь выдвигает гипотезу и предсказывает размер эффекта, полученный на основе количественной теории, которая ранее была опубликована. Он предварительно объявляет эту гипотезу и соответствующий план эксперимента. Впоследствии в эксперименте обнаруживается предсказанное различие. Такое экспериментальное открытие можно интерпретировать как сильный аргумент в пользу гипотезы и количественной теории, но не похоже, что предварительное объявление имеет к такому подтверждению хоть какое-то отношение. Поскольку теория была опубликована ранее, другие исследователи могли бы пойти по стопам автора оригинальной статьи, вывести то же самое предсказание размера эффекта и сделать вывод, что план эксперимента был правильным. В подобной ситуации предварительно регистрировать план эксперимента кажется излишним, поскольку его обоснование вытекает из уже известных идей.

На практике большинство исследований не относится к таким крайним случаям, но ученые часто планируют эксперименты, руководствуясь комбинацией неясных идей, интуиции, любопытства, результатов прошлых экспериментов и количественных теорий. Невозможно измерить качество плана эксперимента для неопределенных составных частей, и предварительное объявление эту ситуацию никак не изменит. Для тех частей предсказанных гипотез (а также методов и измеренных показателей), которые количественно выводятся из существующей теории или знания, качество эксперимента можно измерить по легко доступной информации, и предварительное объявление к этому качеству ничего не добавляет.

Предварительное объявление действительно вынуждает исследователей делать реальное предсказание, а затем ставить эксперимент, который надлежащим образом его проверяет. Это достойная одобрения цель. Но такая цель не имеет смысла, если исследователь не надеется ее достичь. Когда исследователь планирует эксперимент, руководствуясь туманными идеями, он производит предварительную проработку, и довольно глупо просить (и даже настойчиво приглашать) такого исследователя сделать предсказание. Под давлением он может что-то родить, но такие предсказания будут никак не связаны с процессом, в результате которого порождены. В луч-

шем случае подобные исследования дадут какую-то информацию об интуиции ученого, но обычно исследователям не очень интересно, умеют ли другие ученые генерировать хорошие догадки. Они ставят эксперименты, чтобы проверить различные аспекты теоретических заявлений.

На практическом уровне многие исследователи, соблазнившиеся возможностью предварительно объявить о своих гипотезах, могут довольно быстро осознать, что сделать это не получится, потому что их теории недостаточно точны. Быть может, это станет для них полезным открытием и в перспективе может даже послужить улучшению качества научных работ. Кроме того, предварительное объявление действительно имеет отношение к некоторым типам степеней свободы исследователя, как то: необязательная остановка, отбрасывание не приведших к успеху условий и выдвижение гипотез после того, как стали известны результаты. Но это именно те проблемы, которые решаются хорошо обоснованным планом эксперимента.

Короче говоря, формулирование обоснований для плана эксперимента может быть хорошим упражнением для ученых, поскольку открывает возможность самоконтроля. Выписывание всех деталей и обоснований хорошо и потому, что спустя время обоснования забываются. Добавим еще, что попытка быть максимально точным часто приводит к осознанию ученым того факта, что часть его работы изыскательская. А выявление изыскательских частей исследования может стать руководством к интерпретации и презентации эмпирических открытий. Однако обоснование плана эксперимента должно быть частью стандартного научного отчета об эксперименте, поэтому не видно никаких дополнительных преимуществ в публикации обоснования заранее в составе предварительного объявления.

12.3. Альтернативные виды статистического анализа

У критики проверки гипотез долгая история, и зачастую критики предлагают альтернативные виды анализа, которые якобы должны улучшить статистический вывод. Не отрицая проблем, свойственных традиционной проверке гипотез, и лично отдавая предпочтение некоторым альтернативным статистическим методам, мы все же хотим подчеркнуть, что важно понимать, что именно делает метод, и только потом решать, подходит ли он для конкретного научного исследования.

Например, критики проверки гипотез иногда заявляют, что *p*-значения, используемые в этой методологии, бессмысленны, за-

шумлены или ни о чем не говорят. В зависимости от ситуации в таких заявлениях может содержаться зерно истины, но в целом эта точка зрения не может быть истинной, потому что *p*-значение зачастую основано в точности на той же информации, присутствующей в наборе данных (оценка отношения сигнала к шуму), что и другие статистики. Например, когда анализ основан на двухвыборочном *t*-критерии с известными размерами выборок n_1 и n_2, *t*-значение можно преобразовать во много других статистик. По адресу *http://psych.purdue.edu/~gfrancis/EquivalentStatistics/* имеется онлайновое приложение для таких преобразований.

Из этой эквивалентности информации в разных видах статистики следует, что действительно важна не конкретная статистика, а выводы, которые делаются на ее основе. Если вас интересует определенный вывод, то следует использовать соответствующую статистику. Выбор разных статистик может давать совершенно различные ответы, потому что имеются в виду разные вопросы. Без подробных объяснений приведем несколько примеров, иллюстрирующих эту идею (точное значение встречающихся ниже терминов понимать необязательно). Если $n_1 = n_2 = 250$ и $d = 0.183$, то (при выполнении всех прочих условий) все приведенные ниже выводы из данных верны:

- $p = 0.04$, т. е. меньше типичного порога 0.05. Этот результат статистически значим;
- $CI_{95} = (0.007, 0.359)$ – доверительный интервал для *d* Коэна; часто его интерпретируют как степень неуверенности в истинном размере эффекта в генеральной совокупности;
- $\triangle AIC = 2.19$, это разность информационных критериев Акаике для нулевой и альтернативной моделей. Такое значение показывает, что модель с различными средними предсказывает будущие данные лучше, чем модель с одинаковыми средними двух генеральных совокупностей;
- $\triangle BIC = -2.03$, это разность байесовских информационных критериев для нулевой и альтернативной моделей. Такое значение показывает, что нулевая модель верна;
- JZS BF = 0.755, это коэффициент Байеса, основанный на априорном распределении Джеффриса–Целлнера–Сиу, показывающий наличие слабого свидетельства в пользу верности нулевой модели.

Таким образом, эти данные дают значимый результат ($p < 0.05$) и отдают предпочтение альтернативной модели ($\triangle AIC > 0$), но вместе с тем содержат слабое свидетельство в пользу верности нулевой модели ($\triangle BIC < 0$ и JZS BF < 1). Эти выводы могут показаться проти-

воречащими друг другу, но они различны, потому что различаются вопросы. Если вы хотите класть в основу решений об эффектах процесс, контролирующий частоту ошибок типа I, то искомый ответ дает *p*-значение. Если вам нужно получить представление о степени недостоверности оценки стандартизованного размера эффекта, то доверительный интервал – именно то, что требуется. Если требуется оценить, какая модель лучше предсказывает будущие данные – с различными или одинаковыми средними, – то ответ дает значение $\triangle AIC$. Если ваша цель – определить, содержат ли данные свидетельства в пользу нулевой (средние одинаковы) или альтернативной (средние различны) модели, то обратитесь к $\triangle BIC$ или JZS BF. Отметим, что у подходов на основе $\triangle AIC$, $\triangle BIC$ и BF имеется возможность *принять* нулевую гипотезу. Стандартная процедура проверки гипотез никогда не принимает нулевую гипотезу, потому что отсутствие доказательства не есть доказательство отсутствия (глава 3, следствие 3а).

12.4. Роль воспроизводимости

Многие ученые считают воспроизводимость окончательным арбитром в эмпирических вопросах. Если открытие воспроизводится (быть может, независимой лабораторией), то результат считается доказанным. Неудача воспроизведения поднимает вопросы, на которые нужно дать ответ (ошибку могла допустить как одна, так и другая группа). В главах 9–11 объяснено, что это слишком упрощенный взгляд на воспроизводимость в тех случаях, когда используется статистика. Просто в силу случайности выборки даже хорошо организованные исследования реальных эффектов не всегда допускают успешное воспроизведение.

Интуитивные представления о роли воспроизводимости в науке в значительной мере основаны на ее роли в физике. Например, ускорение свободного падения перышка такое же, как молотка, но только в вакууме, где движению не мешает сопротивление воздуха. Вторая часть этого предложения особенно важна, потому что подчеркивает, что результат зависит от условий и особенностей эксперимента. Так, чтобы успешно воспроизвести ускорение свободного падения в вакууме, необходимо точно измерять время и расстояние; таймер с фотозатвором – отличное приспособление для экспериментатора с хронометром. Кроме того, ньютоновская физика постулирует, что не имеет значения, проводится ли эксперимент утром или в полдень, мужчиной или женщиной. Также вполне можно заменить перышко и молоток собачьим кормом и линкором.

Успех воспроизведения в физике почти всегда расценивается относительно некоторой теории. Есть много экспериментальных фактов, подтверждающих правильность ньютоновской физики и независимость ускорения свободного падения от предмета при условии, что необходимые условия соблюдены, а измерения точны. В таких случаях неудачное воспроизведение приобретает особый интерес, потому что свидетельствует о проблеме в постановке эксперимента (быть может, отказал вакуумный насос) или в теории (постоянство скорости фотонов даже под воздействием гравитации привело к теории относительности Эйнштейна). Наука изобилует историями о том, как успешное воспроизведение стало причиной торжества теории (воспроизведение как подтверждение), и о том, как неудачная попытка воспроизведения стала побудительным мотивом для разработки теории.

В отличие от психологии, социологии, биологии и многих других наук у воспроизведения в физике есть важная особенность – результат эксперимента (почти) всегда детерминирован. Много усилий приходится прилагать для идентификации и устранения источников шума. Например, наивный экспериментатор мог бы отпускать предметы из левой и правой руки, что привело бы к случайным различиям в моментах отпускания. При организации эксперимента со свободным падением правильнее было бы использовать механическое устройство, откалиброванное так, чтобы отпускать оба предмета одновременно, устранив тем самым один источник шума. Для многих физических явлений только отсутствие мотивации и ресурсов для устранения неопределенности ограничивает такой вид тщательного контроля.

В экспериментальной психологии, медицине и смежных областях ситуация иная. Какие-то источники шума устранить можно (например, изменив шкалу измерений или научив обследуемых правильным действиям), но внутренне присущие проблеме ограничения часто неустранимы ни при какой мотивации и ресурсах экспериментатора. А еще важнее то, что измеряемый эффект часто обладает естественной изменчивостью (например, одни люди демонстрируют эффект, а другие нет), т. е. изменчивостью, которая является частью явления, а не привнесена шумом (глава 3, следствие 4). В результате у исследователя не остается другого выбора, кроме как использовать статистические методы, например проверку гипотез, а они иногда дают противоречивые результаты просто в силу изменчивости выборки.

Чтобы в таких исследованиях воспроизводимость можно было использовать так же, как в физике, результат должен повторяться почти каждый раз (высокая мощность) и (или) должна существовать

теория, позволяющая различить изменчивость выборки и изменчивость измерения (глава 3, следствие 4а).

12.5. Упор на механизмы

В этой книге мы неоднократно видели, что проблемы статистики возникают во многих науках. И, как описано в последних главах, они объясняются не только неправильным применением на практике, а зачастую носят концептуальный характер. В главе 3, следствия 4, мы видели, что в науках, основанных в первую очередь на статистике, часто невозможно отделить истинную изменчивость от шума, а потому остается открытым вопрос о том, действительно ли существует значимый эффект, или он имеет место лишь для части генеральной совокупности. Кроме того, в предыдущем разделе мы показали, что неудача эксперимента мало о чем говорит. У нас сложилось ощущение, что присущие статистике проблемы так занимают ученых, что они упускают из виду, быть может, самые фундаментальные основы науки: описание и понимание механизмов, которые приводят к наблюдаемому результату. Без такого понимания невозможно предсказать будущие результаты или с уверенностью утверждать, что эмпирическое открытие будет воспроизведено в новых условиях.

Для уверенности в эмпирических результатах необходимо, чтобы соответствующая теория описывала необходимые и достаточные условия. Даже те экспериментальные находки, которые, вообще говоря, воспроизводимы, не могут вселить достаточную уверенность без теоретического объяснения, потому что нет гарантии, что в новых условиях эксперимент даст тот же результат.

Рассмотрим потенциальные различия между двумя магнитно-резонансными томографами (МРТ), показанными на рис. 12.1. Томограф на правом верхнем рисунке находится в Лозанне, Швейцария (левый верхний рисунок), а томограф на правом нижнем рисунке – в Западном Лафайете, штат Индиана (левый нижний рисунок). Откуда инженеры знают, что оба прибора работают одинаково? Между Лозанной и Западным Лафайетом так много различий, которые потенциально могли бы повлиять на поведение прибора. Лозанна расположена рядом с горами, озером, здания там каменные, а жители едят фондю и говорят по-французски. Рядом с Западным Лафайетом раскинулись соевые поля, протекает река, здания кирпичные, а жители едят гамбургеры и говорят по-английски. Откуда мы знаем, что эти различия не влияют на работу приборов? Недостаточно знать, что другие МР-томографы по видимости работают похоже,

ведь каждый новый прибор устанавливается в своем окружении, а проверить все возможные окружения вряд ли получится.

У инженеров есть уверенность в поведении МР-томографов, потому что они понимают, как эти приборы работают. Например, в современном МРТ используются свойства сверхпроводимости, экспериментально открытой в 1911 году. Хотя сверхпроводимость можно было воспроизвести в различных условиях, лишь в 1930-х годах появилась количественная теория, объясняющая это явление. Последующие работы 1930-х годов объяснили сверхпроводимость в терминах куперовских пар и тем самым связали ее с физикой конденсированного состояния и квантовой механикой. Такое понимание позволяет ученым описать необходимые и достаточные условия возникновения сверхпроводимости и предсказать ее свойства в МР-томографах и других устройствах. Инженеры уверены, что томограф будет работать не потому, что так было раньше, а потому что понимание теории сверхпроводимости (и многих других аспектов МРТ) позволяет утверждать, что он будет работать, невзирая на различия в окружении.

Рис. 12.1. Два магнитно-резонансных томографа: в Лозанне, Швейцария (вверху), и в Западном Лафайете, штат Индиана (внизу)

В качестве еще одного примера рассмотрим чуму, унесшую жизни почти трети населения Европы несколько веков назад. В 1898 году французский ученый Поль-Луи Симон установил, что переносчиками чумы являются блохи, паразитирующие на крысах. Тщательно поставленные эксперименты обосновали этот механизм (он также выявил бактерию *Yersinia pestis*, заражавшую блох), поскольку было показано, что блохи, перепрыгивающие с зараженной крысы на здоровую, переносили чуму. Связь между крысами и чумой является (неполным) механизмом: крысы несут чуму. Из этого механизма вытекает ясное следствие: чтобы уменьшить распространение чумы, нужно сократить поголовье крыс: держите в доме собак и кошек, охотящихся на крыс, храните еду в герметичной упаковке, ставьте крысоловки, избегайте контактов с живыми или дохлыми крысами (как это ни печально, подозревали, что одним из механизмов распространения Великой Лондонской чумы в 1665 году были собаки и кошки, поэтому их уничтожали во множестве, что, возможно, привело к увеличению популяции крыс). Когда в 1900 году в Сан-Франциско появилась чума, ученые рекомендовали убивать крыс (но из-за политического конфликта этот добрый совет не удалось претворить в жизнь). Отметим, что механизм связи между чумой и крысами не обязан давать количественные предсказания о том, сколько жизней будет спасено в результате тех или иных действий; он говорит лишь, что почти любое действие, направленное на уменьшение контактов с крысами, поможет обуздать чуму. Он также указывает, какие действия, вероятно, окажутся бесполезными (например, изоляция зараженных домочадцев). Конечно, теории, включающие более точные механизмы, лучше. В наши дни чуму лечат антибиотиками, которые непосредственно атакуют бактерии, являющиеся первопричиной.

И напоследок рассмотрим эффект анестезии. Действие изофлурана (фторированного простого эфира), одного из анестетиков, известно уже больше ста лет; благодаря ему стали возможны сложные хирургические операции. И хотя эффект очень надежен, механизм анестезии изофлураном до сих пор неизвестен. Более того, зарегистрированы случаи, когда пациент просыпался в середине операции или сохранял воспоминания об операции. Если бы мы понимали механизм анестезии изофлураном, то могли бы предвидеть и предотвращать такие случаи. Не имея теории, мы не знаем, является ли неудача шумом или признаком некоего скрытого механизма. А тем временем врачи применяют анестезию, понимая, что должны быть готовы к неудаче.

Идентифицировать механизмы и обосновать их роль в наблюдаемом явлении очень трудно, но это должно быть долгосрочной целью каждого ученого. Некоторые так никогда и не достигают этой цели, поскольку тратят свое время на сбор данных и проверку идей (тоже очень полезное дело) – деятельность, которая может помочь будущим ученым идентифицировать и обосновать механизмы. Пока механизмы не получат теоретического обоснования, ученые не могут быть уверены в том, что некоторый эффект будет иметь место и в новых условиях.

В некотором смысле долгосрочной целью науки является устранение (в той мере, в какой это возможно) статистики путем отыскания «правильных» факторов и уменьшения изменчивости (см. главу 6). Понимание механизмов позволяет выдвигать новые гипотезы, которые можно строго исследовать в правильно спланированных экспериментах. Например, понимание того, как витамины воздействуют на различные органы, объяснило бы, почему витамины благотворно сказываются на здоровье одних людей и вредят другим. Таким образом, благодаря более глубокому постижению механизмов многие проблемы и опасения, о которых шла речь в этой книге, просто исчезли бы. Таким образом, если мы хотим вдохнуть новые силы в научную практику, нашей целью должно быть не реформирование статистики, а устранение нужды в ней.

Что следует запомнить

1. Многие предложения по улучшению практики применения статистики, например предварительное объявление, не решают фундаментальных проблем.
2. Хорошая наука отнюдь не сводится к статистике.

Книги издательства «ДМК Пресс» можно заказать
в торгово-издательском холдинге «КТК Галактика» наложенным платежом,
выслав открытку или письмо по почтовому адресу:
115487, г. Москва, пр. Андропова д. 38 оф. 10.
При оформлении заказа следует указать адрес (полностью),
по которому должны быть высланы книги;
фамилию, имя и отчество получателя.
Желательно также указать свой телефон и электронный адрес.
Эти книги вы можете заказать и в интернет-магазине: **www.galaktika-dmk.com**.
Оптовые закупки: тел. **(499) 782-38-89**.
Электронный адрес: **books@alians-kniga.ru**.

Майкл Х. Херцог, Грегори Фрэнсис, Аарон Кларк

Статистика и планирование эксперимента для непосвященных

Главный редактор *Мовчан Д. А.*
dmkpress@gmail.com
Зам. главного редактора *Сенченкова Е. А.*
Корректор *Абросимова Л. А.*
Верстка *Луценко С. В.*
Дизайн обложки *Мовчан А. Г.*

Формат 60×90 1/16.
Гарнитура «PT Serif». Печать цифровая.
Усл. печ. л. 10,88. Тираж 100 экз.

Веб-сайт издательства: *www.dmkpress.com*