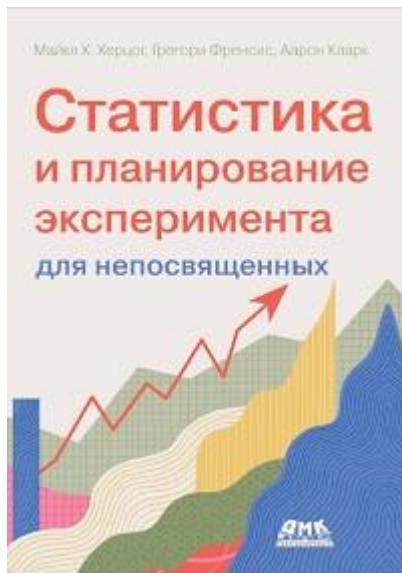


Майкл Херцог. Статистика и планирование эксперимента для непосвященных

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, 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.

Майкл Х. Херцог, Грэгори Френсис, Аарон Кларк. Статистика и планирование эксперимента для непосвященных: Как отучить статистику лгать – М.: ДМК Пресс, 2023. – 174 с.¹



Купить книгу в [Ozon](#) или [Лабиринте](#)

Часть I. Принципы статистики

Глава 1. Основы теории вероятностей

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.

Отношении шансов (ОШ) Многие курильщики умирают от инфаркта. Бросать курить? Это частичная информация! Встречный вопрос: сколько некурящих умирает от инфаркта?

Гипотетический пример: из 107 курящих семь перенесли инфаркт, т. е. 100 не переносили (рис. 1). Шансом называется отношение 7/100. Для некурящих инфаркт перенес 1 человек из 101, поэтому шанс равен 1/100. Идея отношения шансов (ОШ) – сравнить две дроби, поделив одну на другую. Это отношение двух дробей говорит нам, в какой степени курящие страдают от инфаркта чаще, чем

¹ Обычно книги издательства ДМК Пресс отличает хорошее качество. Здесь же количество опечаток зашкаливает. Благо оригинальное издание на английском языке находится в открытом [доступе](#). Так что, при подготовке конспекта я постоянно сверялся с оригиналами. – Прим. Багузина

некурящие: $(7/100) / (1/100) = 7$. Таким образом, у курящего шанс получить инфаркт в семь раз больше, чем у некурящего, – немало.

A	Куря- щие	Неку- рящие	B	Курящие	Некуря- щие
Инфаркт был	7	1	Инфаркт был	7	1
Инфаркта не было	100	100	Инфаркта не было	10 000	10 000

Рис. 1. Курение и смерть от инфаркта

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!

Глава 2. Планирование эксперимента и основы статистики: теория обнаружения сигналов (ТОС)

Классический сценарий ТОС

Мы находимся в подводной лодке. Гидролокатор излучает волны, а приемник получает их отражения. Эти отраженные волны объединяются в гидроакустическую характеристику.

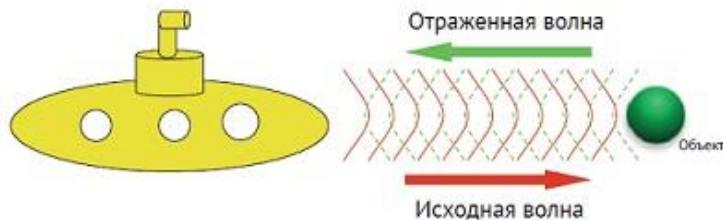


Рис. 2. Подлодка посылает гидроакустические сигналы и измеряет отраженные сигналы

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. Каждому из двух возможных условий соответствует распределение вероятностей, показывающее, насколько вероятно некоторое значение гидроакустической характеристики на оси x.

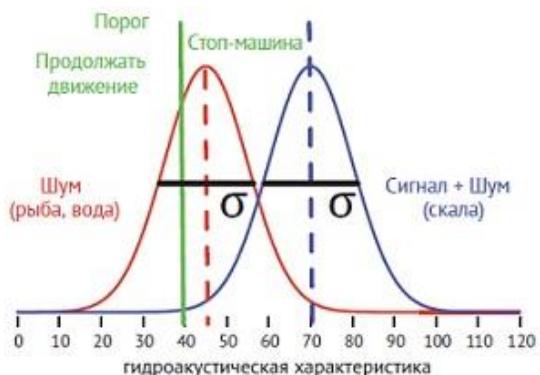


Рис. 3. Классический сценарий ТОС. Оба условия – присутствие и отсутствие скалы – принадлежат какому-то распределению вероятностей, которое показывает, насколько вероятно получение различных значений гидроакустической характеристики

Мы предполагаем, что ситуацию описывает нормальное распределение. Оно полностью задается 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 будут значительно различаться. Если $\sigma = 0$, 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, а присутствие скалы – сигнал плюс шум.

Насколько хорошо мы можем отличить присутствие скалы от отсутствия? Зависит от перекрытия гауссиан. Перекрытие можно оценить разностью между средними гауссиан, поделенной на стандартное отклонение, в предположении, что σ для обоих распределений одинаково:

$$(1) d' = \frac{\mu_1 - \mu_2}{\sigma}$$

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).

ТОС и доля правильных ответов

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

Рис. 4. Четыре исхода в классическом эксперименте ТОС

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): $p = (Hit + CR) / 2$.

К сожалению, процент верных ответов смешивает порог принятия решения и различимость. 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 способен затушевывать истинные эффекты или давать ложноположительные результаты. Выводы, базирующиеся на проценте верных ответов, основаны на частичной информации!

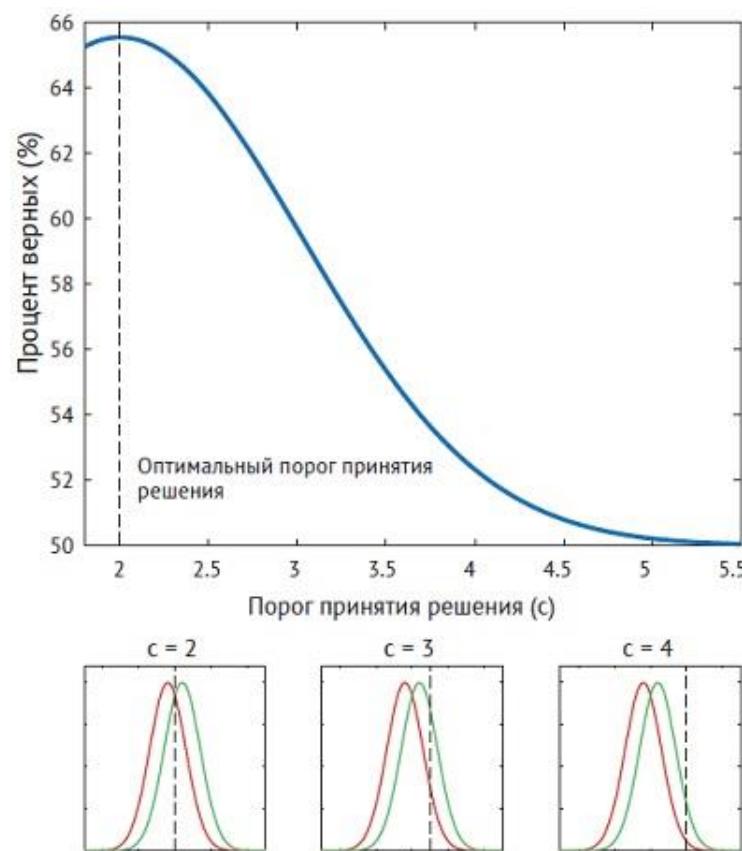


Рис. 5. Процент верных ответов зависит от порога. Сначала поместим порог в точку пересечения двух гауссиан, т. е. положим равным 2 (левый нижний рисунок). Затем будем сдвигать порог вправо

(средний и правый рисунок). На верхнем рисунке показано, что процент правильных ответов уменьшается по мере отодвигания порога от оптимума.

Эмпирическая d'

Мы можем разделить различимость (d') и смещение порога (b), если будем оценивать их по отдельности:

$$(2) d'_{emp} = z(Hit) - z(FA) \quad b_{emp} = -\frac{z(Hit) + z(FA)}{2}$$

Здесь z – обратная гауссова функция распределения.²

Сравним работу врача и системы на основе искусственного интеллекта (ИИ) при диагностике заболевания. На рис. 6 приведены частоты правильных подтверждений, ложных тревог, ошибочных пропусков и правильных пропусков.

Качество работы врача			Автоматическое распознавание		
Сигнал	Есть	Нет	Сигнал	Есть	Нет
Да	80	20	Да	98	38
Нет	20	80	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. 6. Сравнение врача с искусственным интеллектом

Процент правильных ответов у врача = $(80+80)/2 = 80$ и ИИ = $(98+62)/2 = 80$ одинаков. А вот чувствительность d' различается. Обычно чувствительность – встроенная характеристика системы, изменить которую трудно. Изменить порог принятия решения проще. ИИ имеет сильное смещение в сторону ответов «да», что помогает сократить ошибочные пропуски (ложноотрицательные результаты, Miss у врача = 0,2, у ИИ = 0,02), но увеличивает частоту ложных тревог (ложноположительных результатов, FA у врача = 0,2, у ИИ = 0,38). Порог ИИ далек от оптимального.

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).

Глава 3. Главная концепция статистики

Обычно одну из генеральных совокупностей мы называем чистым шумом со средним μ_N , а другую – зашумленным сигналом со средним μ_{SN} . Часто у нас нет информации о генеральной совокупности, но мы можем провести серию измерений и вычислить выборочные средние \bar{x} , стандартное отклонение s и оценить размер эффекта, называемый d Коэна:

$$(3) d = \frac{\bar{x}_{SN} - \bar{x}_N}{s}$$

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

- Частота правильных подтверждений = Мощность.
- Частота ложноположительных результатов = Ложная тревога = Ошибка I-го рода.
- Частота ошибочных пропусков = Ошибка II-го рода.
- d Коэна = Размер эффекта.

² В Excel для её вычисления используется функция НОРМ.СТ.ОБР(вероятность). В нашем примере $z(Hit)$ у врача = НОРМ.СТ.ОБР(0,8) = 0,842, у ИИ = НОРМ.СТ.ОБР(0,98) = 2,054.

Недостаточная выборка

Если мы сделали выборку из генеральной совокупности и подсчитали выборочное среднее \bar{x} , то оценку того, насколько выборочное среднее может отличаться от истинного математического ожидания генеральной совокупности μ , дает *стандартная ошибка среднего*:

$$(4) \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

где σ – стандартное отклонение распределения генеральной совокупности, n – размер выборки. Поскольку σ не известно, мы можем оценить стандартную ошибку среднего, рассчитав её выборочное значение:

$$(5) s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

При увеличении размера выборки выборочная ошибка уменьшается. С другой стороны, из-за недостаточности выборки выборочное среднее, может серьезно отличаться от истинного среднего.

Сравнение средних

Допустим мы хотим сравнить высоту дубов на северном и южном склоне Альп. Мы взяли две выборки и рассчитали их среднее. Они не равны. 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? Вычислив d Коэна для альтернатив.

Для первой альтернативы – высота дубов на двух склонах одинаковая – $\mu_{North} - \mu_{South} = 0$. Это значит, что мы имеем распределение чистого шума. Соответствующее выборочное распределение центрировано вокруг 0. Для второй альтернативы различие присутствует, и выборочное распределение центрировано относительно $\mu_{North} - \mu_{South}$. Поскольку истинные значения неизвестны, мы используем оценки.

Итак, мы оценили два выборочных распределения: в одном налицо различие (сигнал и шум) со средним $\mu_{North} - \mu_{South}$, а в другом различия нет (только шум) и среднее равно нулю. Оценка d Коэна, которая для случая выборочных средних называется t , может быть рассчитана по формуле (3). В этом случае формула принимает вид:

$$(6) t = \frac{(\bar{x}_{SN} - \bar{x}_N) - (0)}{s_{\bar{x}_{North} - \bar{x}_{South}}}$$

Оценка стандартной ошибки разности средних:

$$(7) s_{\bar{x}_{North} - \bar{x}_{South}} = \sqrt{\frac{s_{\bar{x}_{North}}^2}{n_{North}} + \frac{s_{\bar{x}_{South}}^2}{n_{South}}} \text{ Если } s_{\bar{x}_{North}} = s_{\bar{x}_{South}}, n_{North} = n_{South}, \text{ то } s_{\bar{x}_{North} - \bar{x}_{South}} = s \sqrt{\frac{2}{n}}$$

Стандартная ошибка объединяет два источника неопределенности: шум s и размер выборки n . Подставим оценку стандартной ошибки разности средних в уравнение (6):

$$(8) t = \frac{(\bar{x}_{SN} - \bar{x}_N) - (0)}{s_{\bar{x}_{North} - \bar{x}_{South}}} = \frac{(\bar{x}_{SN} - \bar{x}_N) - (0)}{s \sqrt{\frac{2}{n}}} = \frac{\bar{x}_{SN} - \bar{x}_N}{s} \cdot \sqrt{\frac{n}{2}} = d \cdot \sqrt{\frac{n}{2}}$$

t -значение объединяет размер эффекта d с размером выборки n .

Ошибки I-го и II-го рода

Недостаточность выборки может стать причиной ошибки I-го рода, 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 I-го рода, а ошибочный пропуск – ошибкой II-го рода.

Ложные тревоги и ошибочные пропуски зависят от заданного нами порога. Возможные исходы показаны на рис. 7. 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_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. 7 эта ситуация обозначена фразой « H_0 неверна».



Рис. 7. Задача статистики – делать заключения о гипотезе. (1) Нулевая гипотеза неверна, средние различны (Правильное подтверждение). (2) Нулевая гипотеза верна, но мы пришли к выводу, что средние различны (Ложная тревога, или ошибка I-го рода). (3) Нулевая гипотеза неверна, но у нас недостаточно фактов, свидетельствующих против нее (Ошибкаочный пропуск, или ошибка II-го рода). (4) Нулевая гипотеза верна, и у нас недостаточно фактов, свидетельствующих против нее (Правильный пропуск)

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.

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 таких областях, как медицина, психология, нейронауки и биология, обычно применяется порог, который в первом приближении следует «правилу двух сигм». При этом стоимость получения одного образца для медицинской или биологической выборки часто гораздо выше, чем в физике.

ТОС говорит, что какой порог ни выбрать, имеет место компромисс между правильными подтверждениями и ложными тревогами.

Вместо того чтобы задавать порог в терминах стандартного отклонения σ , во многих областях (включая медицину, психологию, нейронауки и биологию) исследователь хочет, чтобы частота ошибок I-го рода 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.

Ошибка I-го рода: р-значение связано с порогом

Распределение разности выборочных средних, когда нулевая гипотеза H_0 верна, 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 совершаляем ошибку I-го рода. Вероятность t -значения большего $t_{cv} = 2,0$ равна площади под кривой в области правее $t_{cv} = 2,0$.

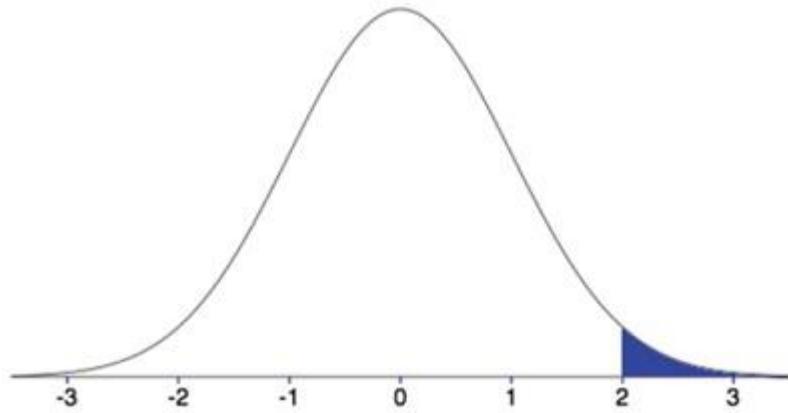


Рис. 8. Связь между критическим значением порога и частотой ошибок I-го рода. Кривая показывает распределение чистого шума, т. е. случай, когда нулевая гипотеза верна.

Такой тест называется «односторонним t -критерием». Площадь закрашенной области зависит от размера выборки. Например, для $n = 20$, площадь (она же вероятность) в Excel можно вычислить с помощью формулы =СТЬЮДЕНТ.РСП(2;20) = 0,0296. Следовательно, при использовании порога $t_{cv} = 2,0$ мы допустим ошибку I-го рода с вероятностью 2,96%. Еще раз, в чем ошибка? Мы отклоняем нулевую гипотезу о равенстве средних, хотя существует вероятность равная 2,96%, что различие между выборочными средними обусловлено только шумом.

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 р-значением. Оно говорит, насколько вероятно при условии истинности нулевой гипотезы получить наше или даже большее t -значение.

Если р-значение меньше 0,05, то эффект называется значимым. Следовательно, для контроля частоты ошибок I-го рода нужно лишь потребовать, чтобы вычисленное р-значение было меньше желаемой частоты ошибок.

Поскольку р-значение определяется t -значением, оно смешивает в одну кучу размера эффекта (d) и размера 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 р-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.

	Малый	Средний	Большой
Размер эффекта	0.2	0.5	0.8

Рис. 9. Рекомендации Коэна по размеру эффекта d

Следствия, комментарии и парадоксы

Еще раз подчеркнем, что t -значение, а стало быть, и р-значение, определяются оценочным d и размером выборки n :

$$(9) t = d \cdot \sqrt{\frac{n}{2}}$$

Если оценочное значение $d \neq 0$, 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.

Если оценочное значение $d \neq 0$, то существует размер выборки $n < m$ такой, что t -критерий не является значимым для n , но является значимым для m . 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. Когда мы не отвергаем нулевую гипотезу нельзя сделать никаких выводов. Это ключевая проблема проверки гипотез.

Само по себе p -значение ничего не говорит о размере эффекта. Если размер эффекта d остается неизменным, то p -значение уменьшается с увеличением размера выборки.

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 .

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:

$$(9') x_j = \mu + \varepsilon_j$$

где x_j – результат j -го измерения, μ – истинный сигнал, а ε_j – шум, зависящий от испытания. Обычно предполагается, что ε_j имеет нормальное распределение с нулевым средним.

Модель такого вида пригодна во многих областях, в частности в физике. Однако в биологии, медицине и других науках ситуация бывает иной. Зависящие от человека эффекты можно описать следующей формулой:

$$(9'') x_{ij} = \mu + v_i + \varepsilon_{ij}$$

где x_{ij} – одно измерение для пациента i в день j , μ – среднее значение по всей генеральной совокупности, а v_i – чувствительность пациента i к данному болеутоляющему. Таким образом, v_i определяет, насколько один человек отличается от других – и от среднего μ . ε_{ij} – шум измерений, он отражает различие в воздействии лекарства на одного и того же человека в разные дни. В некотором смысле ε_{ij} улавливает несистематическую изменчивость, а v_i – систематическую.

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. Когда v_i отлично от нуля, значимые результаты не позволяют делать заключения на уровне индивидуума.

Глава 4. Вариации на тему t -критерия

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, вегетарианцы и мясоеды.

Типы переменных и шкалы измерений: *номинальная* (страницы), *порядковая* (воинские звания), *интервальная* (значения можно складывать и вычитать, но умножение и деление не имеют смысла;

например, температура в градусах Цельсия), *относительная* (можно умножать и делить, например, вес или рост).

Типы критериев. *Параметрический критерий*, в котором предполагается некоторая модель распределения данных. Например, высоты деревьев в генеральной совокупности распределены нормально. Параметрические распределения можно описать небольшим числом параметров (например, средним и стандартным отклонением). *Непараметрический критерий*: никакое конкретное распределение не предполагается.

Стандартный подход: проверка нулевой гипотезы

Шаги принятия статистического решения в двухвыборочном t-критерии:

1. Сформулировать альтернативную гипотезу, обозначаемую H_1 , например Терапия А отличается от Терапии В.
2. Предположить, что верна гипотеза H_0 : между Терапией А и Терапией В нет различий.
3. Вычислить статистику t-критерия:

$$(10) t = \frac{\bar{x}_A - \bar{x}_B}{s_{\bar{x}_A - \bar{x}_B}} = \frac{\bar{x}_A - \bar{x}_B}{s \sqrt{2/n}}$$

4. Рассчитать p-значение, соответствующее t-критерию. p-значение – площадь под кривой распределения, лежащая за пределами t-критерия (см. рис. 8).
5. Принять решение. Если $p \leq 0,05$, отвергнуть гипотезу H_0 и принять H_1 : 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 I-го рода (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 в 5% случаев. Можно усилить ограничение и потребовать, чтобы $p \leq 0,01$, в таком случае p будет меньше 0,01 только в 1% случаев. Конечно, не обойтись без компромисса: чем строже ограничение, тем больше частота ошибок II-го рода (ошибочных пропусков).

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

$$(11) s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

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

$$(12) t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

а количество степеней свободы $df = n - 1$.

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.

Предположения, лежащие 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.

- Данные должны быть независимы и одинаково распределены
- Распределения генеральной совокупности нормальные

- Поскольку при вычислении t -критерия используются выборочное среднее и выборочное стандартное отклонение, ни номинальные, ни порядковые данные с его помощью анализировать нельзя.
- Стандартный двухвыборочный t -критерий предполагает, что дисперсия обеих генеральных совокупностей одинакова.

Непараметрический подход

Если распределение данных не нормальное, то можно применить непараметрический критерий:

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

Рис. 10. Параметрические критерии и соответствующие им непараметрические критерии

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

Часть II. Множественная проверка гипотез

Глава 5. Задача множественной проверки гипотез

Вычисляя один t -критерий, мы знаем, что если нулевая гипотеза в действительности верна, то частота ошибок I-го рода равна $\alpha = 0,05$. Что то же самое, ошибка I-го рода (ложная тревога) не возникает в 1–0,05% случаях, если нулевая гипотеза верна. При вычислении двух независимых t -критериев вероятность не допустить ложную тревогу равна $0,95^2 = 0,9025$. Для 12 сравнений $0,95^{12} = 0,54$. Таким образом, вероятность хотя бы одной ложной тревоги при 12 сравнениях равна $1 - 0,54 = 0,46$. С увеличением числа сравнений ложные тревоги становятся все более и более вероятными:

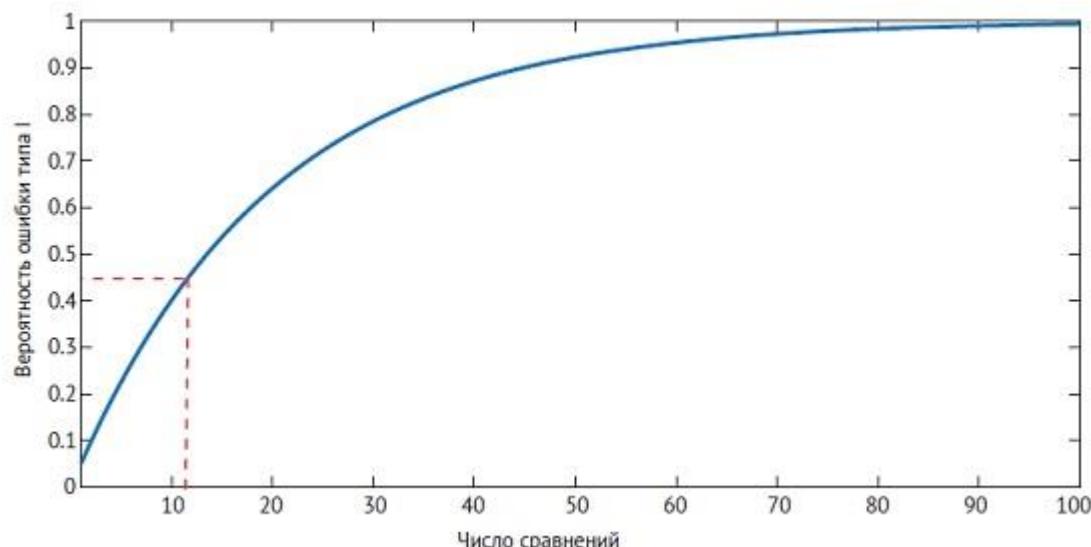


Рис. 11. Частота ошибок I-го рода (частота ложных тревог) сильно зависит от числа сравнений

Поправки Бонферрони. Классический способ учесть увеличение частоты ошибок I-го рода – уменьшить требуемый уровень значимости. Если мы хотим, чтобы частота ошибок I-го рода для m независимых критериев была равна 0,05 для одного сравнения нужно установить $\alpha = 0,05/m$.

Статистики спорят по поводу того, полезны ли поправки Бонферрони (и аналогичные им), и если да, то в каких случаях.

Глава 6. Дисперсионный анализ (ANOVA)

Если необходимо сравнить несколько средних вместо попарного сравнения лучше применить дисперсионный анализ (analysis of variance, ANOVA). Логика ANOVA проста. Мы упрощаем альтернативную гипотезу, считая, что по крайней мере одна из генеральных совокупностей не равна

другим. Например, для сравнения трех выборок высоты деревьев в разных географических зонах мы формулируем одну альтернативную гипотезу вместо трех.

В ANOVA вычисляется отношение двух оценочных дисперсий, которое называется F-значением:

$$(13) F = \frac{\text{Оценка дисперсии на основе изменчивости межгрупповых средних}}{\text{Оценка дисперсии на основе изменчивости внутригрупповых средних}}$$

Формально это выражение можно записать так:

$$(14) F = \frac{\sum_{j=1}^k n_j (M_j - M_G)^2}{\frac{k-1}{\sum_{j=1}^{n_j} \sum_{i=1}^k (x_{ij} - M_j)^2 / (n_j - 1)}}$$

где k – число групп (три генеральные совокупности), 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, x_{ij} – i -й пример в группе j (высота одного дерева). 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.

Если нулевая гипотеза верна (рис. 12а), дисперсии в числителе и в знаменателе близки, а F-значение ≈ 1 . Когда различия между высотами деревьев в трех географических регионах велики (рис. 12б), а σ мало, т. е. высоты деревьев в трех генеральных совокупностях сильно различаются, но внутри одной генеральной совокупности почти равны, изменчивость в основном определяется различиями между группами, и F-значение велико. Обычно ситуация находится 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

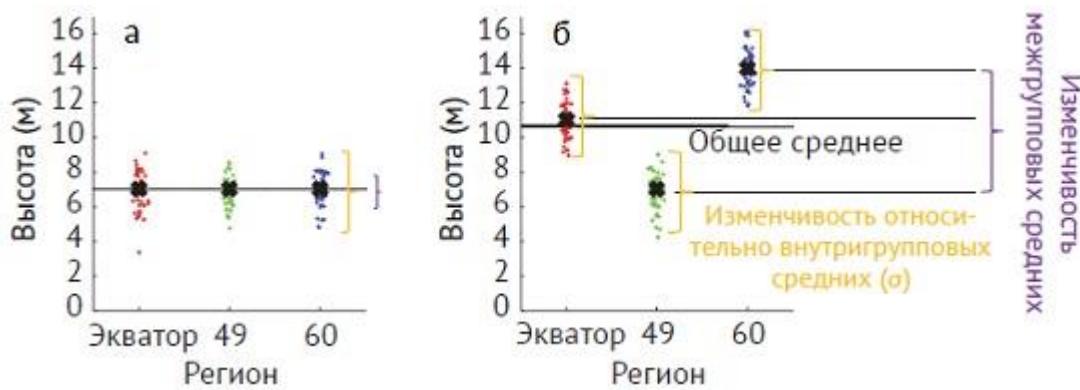


Рис. 12. Логика ANOVA

Как и в случае t-критерия, порог статистической значимости выбран так, чтобы частота ошибок I-го рода 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).

Частным случаем однофакторного ANOVA с независимыми переменными является вариант, когда сравниваются два региона (уровня), как в t-критерии. На самом деле имеется тесная связь между обоими критериями, и в данном случае $F = t^2$. Здесь р-значение будет одинаковым 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 р-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. Полное число степеней свободы $df_1 + df_2 = n - 1$.

Пример вычисления для однофакторного ANOVA с независимыми переменными

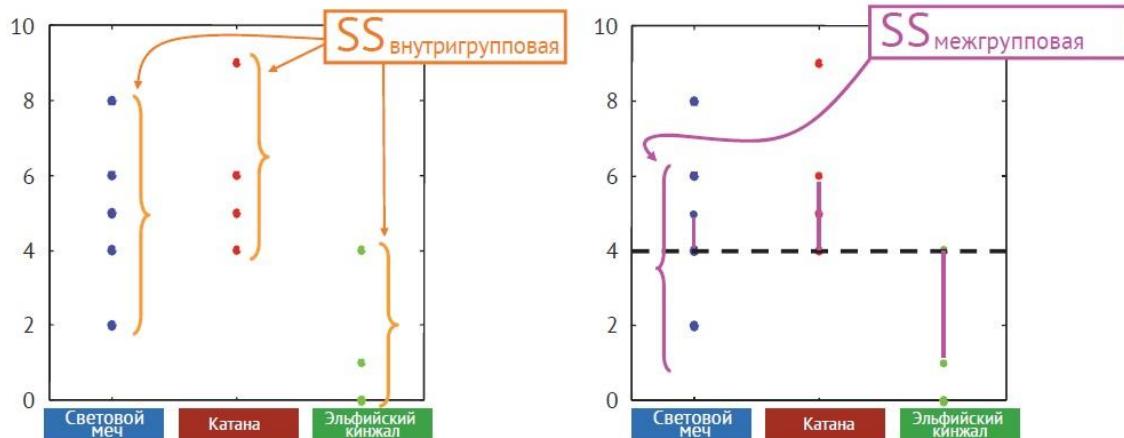
Предположим, что проводится турнир по боям на холодном оружии с тремя разными типами оружия: световым мечом, катаной Хатори Ханзо и эльфийским кинжалом (рис. 13). Вопрос: имеются ли различия в числе побед разным оружием? Нулевая гипотеза заключается в том, что различий нет.

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 вычисления.

Световой меч	$(x_i - M)^2$	Катана	$(x_i - M)^2$	Эльфийский кинжал	$(x_i - M)^2$
6	$(6 - 5)^2 = 1$	6	0	0	1
8	9	5	1	4	9
5	0	9	9	0	1
4	1	4	1	1	0
2	9	6	0	0	1
$M = 5$	$SS = 20$	$M = 6$	$SS = 14$	$M = 1$	$SS = 12$

Общее среднее

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



$$SS_{within} = \sum_{k=1}^3 SS_k = 20 + 14 + 12 = 46$$

$$SS_{between} = \sum_{k=1}^3 n_k (M_k - M_G)^2 = 5(5 - 4)^2 + 5(6 - 4)^2 + 5(1 - 4)^2 = 70$$

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

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

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

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

Рис. 13. Пример вычисления для однофакторного ANOVA с независимыми переменными

F-значение равно 9,14. Это означает, что межгрупповая изменчивость в 9,14 раз больше внутригрупповой изменчивости. Следовательно, большая часть изменчивости проистекает из различия средних, а гораздо меньшая – из изменчивости в пределах каждой группы. F-значение 9,14 приводит к p-значению $0,0039 \ll 0,05$, и мы заключаем, что результаты значимы, т. е. отвергаем

нулевую гипотезу о том, что среднее число побед разным оружием одинаково. 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.

Апостериорные критерии

Идея критерия Шеффе – выполнить несколько сравнений путем вычисления попарных ANOVA. Одно из предположений ANOVA – дисперсии всех генеральных совокупностей одинаковы. Тогда лучшей оценкой изменчивости внутри каждой генеральной совокупности будет объединенная оценка общего ANOVA, в данном случае $MS_{within} = 3,83$.



$$\left. \begin{array}{l} df_{between} = 2 \\ MS_{within} = 3.83 \end{array} \right\} \text{Из рассмотренной выше таблицы ANOVA}$$

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

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

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

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. 14. Вычисление сравнений в апостериорном критерии Шеффе

Сначала вычисляется общее среднее для каждого сравнения (G_{comp}), равное средней сумме всех примеров из рассматриваемой пары групп. Затем вычисляем сумму квадратов отклонений этого среднего от групповых средних по двум группам ($SS_{between}$). Далее находим $MS_{between}$ и F-значение.

р-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 на рис. 15. Между световыми мечами и катанами не удалось найти значимого различия (р-значение = 0,728 выше критического порога $\alpha = 0,05$). А вот, и световые мечи, и катаны отличаются от эльфийских кинжалов (их р-значения 0,023 и 0,006 ниже критического порога $\alpha = 0,05$).

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. 15. Результаты апостериорного критерия Шеффе для трех сравнений

Такие различия часто иллюстрируются с помощью графика [ящик с усами](#).

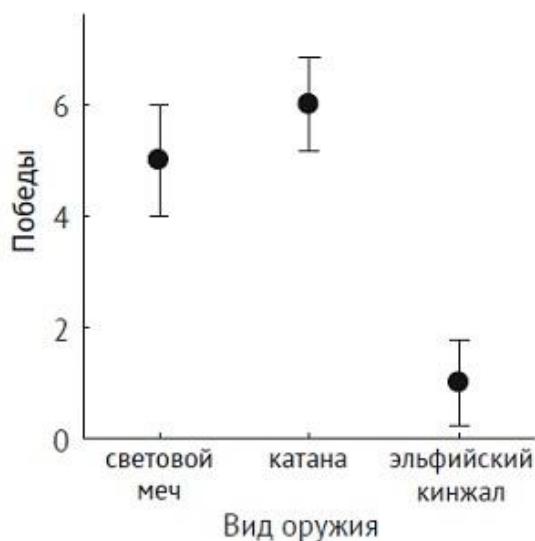


Рис. 16. Среднее число выигрышей и стандартная ошибка для всех типов оружия

Размер эффекта

Как и в случае t-критерия, р-значение в ANOVA смешинает 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:

$$(15) \eta^2 = \frac{SS_{between}}{SS_{total}},$$

где

$$(15.1) SS_{between} = \sum_{j=1}^k n_j (\bar{x}_j - M_G)^2$$

$$(15.2) SS_{total} = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - M_G)^2$$

где M_G – общее среднее по 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 (рис. 17) такой эффект считается большим.

	Малый	Средний	Большой
Размер эффекта	0.01	0.09	0.25

Рис. 17. Рекомендации Коэна по оцениванию размера эффекта

Глава 7. Планирование эксперимента: подгонка модели, мощность и сложные планы

Вероятность успеха эксперимента мы будем оценивать мощностью. Мощность – это частота правильных подтверждений. Иначе говоря, это вероятность того, что случайная выборка позволит

правильно отвергнуть нулевую гипотезу. Если оценить размер эффекта на генеральной совокупности сложно, ориентируйтесь на тот минимальный размер эффекта, который вам интересен. После этого воспользуйтесь компьютерными программами, которые вычисляют мощность.

На рис. 18 показан итог работы бесплатной программы [G*Power](#) при следующих установках:

- двусторонний t-критерий для разности между двумя независимыми средними
- размер эффекта на генеральной совокупности $d = 0,55$
- апостериорный тип анализа мощности
- частота ошибок I-го рода $\alpha = 0,05$
- размеры выборок $n_1 = n_2 = 40$

На графике показаны выборочные распределения для нулевой (красная кривая) и конкретной альтернативной гипотезой (синяя кривая). Синим цветом закрашена область β ошибки II-го рода. Мощность равна $1 - \beta = 0,68$. Т.е., при заданных условиях значимый результат будет получен с вероятностью 0,68.

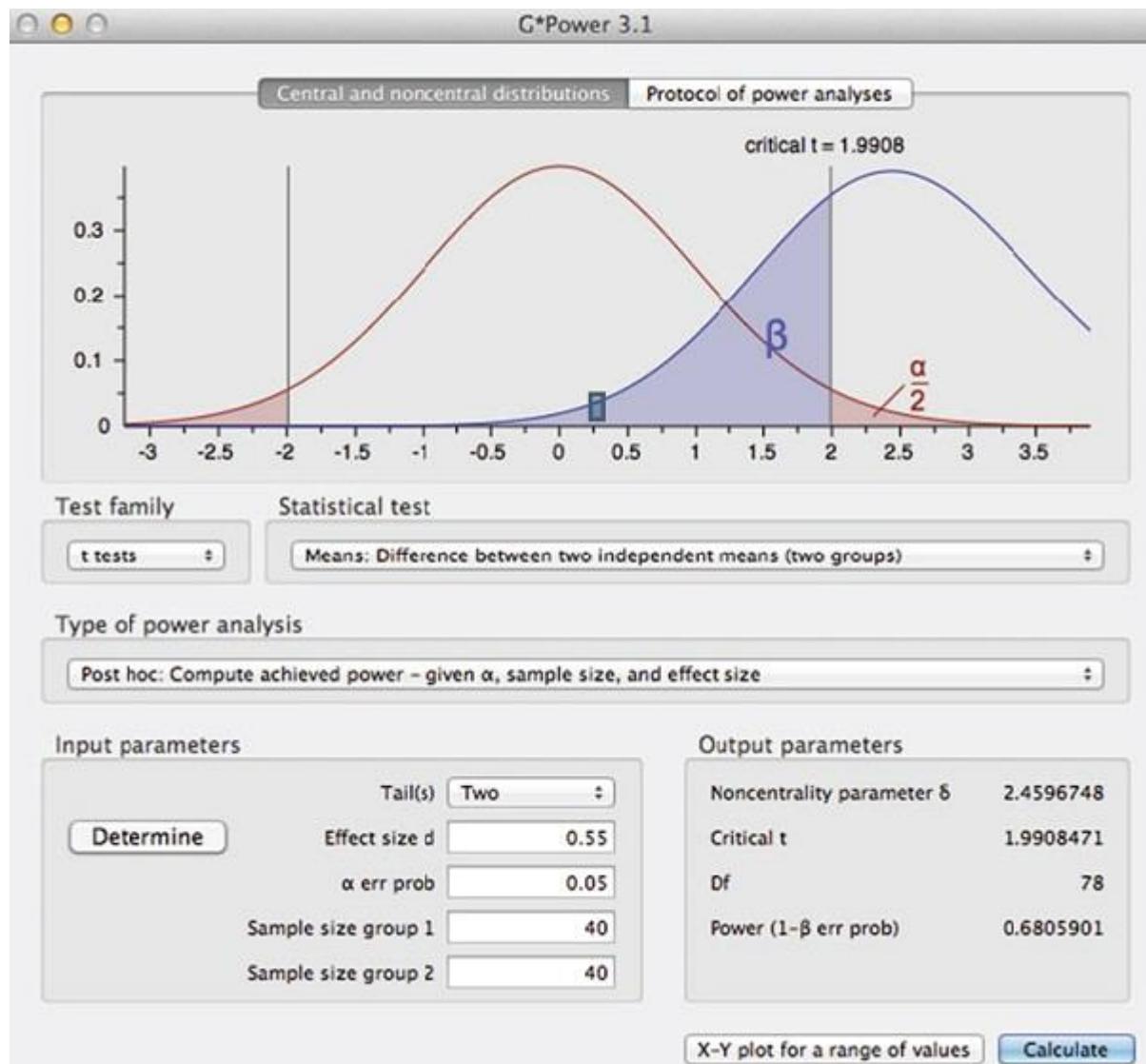


Рис. 18. Вывод программы G*Power, вычисляющей мощность t-критерия

Предположим, что мощность 0,68 нас не устраивает, и мы хотим найти размеры выборок, при которых шанс отвергнуть нулевую гипотезу составляет 90%. В списке Type of power analysis (Тип анализа мощности) выберем *A priori*, а в разделе входных параметров изменим значение поля Power с 0,68 на 0,9. На рис. 19 показан вывод программы в новой ситуации. В разделе выходных параметров мы видим, что для получения мощности 0,9 для двустороннего t-критерия необходимы размеры выборок $n_1 = n_2 = 71$.

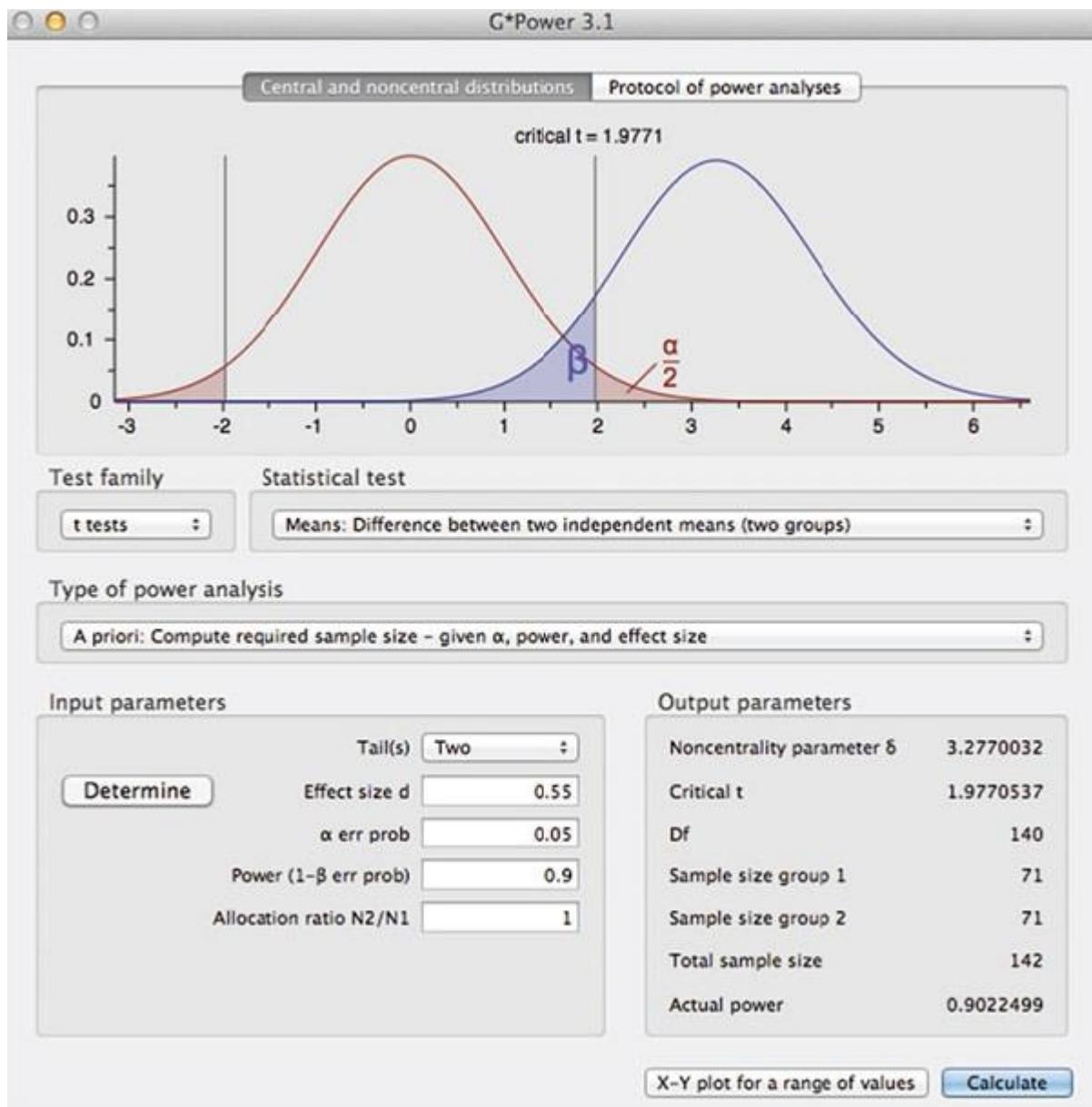


Рис. 19. Вывод программы G*Power, вычисляющей размеры выборок при заданной мощности 0,9

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 смысла 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.

Глава 8. Корреляция

В главе 6 мы изучали влияние географической широты на высоту деревьев, для чего обмеряли деревья в трех местах и проверяли, различаются ли средние высоты. 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. Корреляция позволяет учсть относительность шкалы и тем самым выразить влияние широты одним значением, r .

Корреляция измеряет степень линейной зависимости между двумя переменными. Эта линейная связь описывается формулой ковариации:

$$(16) \text{ cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{X}) \times (y_i - \bar{Y})}{n-1}$$

где x_i – географическая широта, y_i – высота деревьев на этой широте, \bar{X} – средняя широта, \bar{Y} – средняя высота деревьев. Данные представляют собой n пар (широта + высота дерева). Ковариация обобщает понятие дисперсии, потому что $\text{cov}(x, x)$ – 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:

$$(17) r = \frac{\text{cov}(x, y)}{s_x \cdot s_y}$$

Корреляция этого типа называется коэффициентом корреляции Пирсона. Его значения изменяются от -1.0 до +1.0, где -1 означает отрицательную линейную зависимость, 0 – отсутствие корреляции, +1 – положительную линейную зависимость.

Проверка гипотез с помощью корреляции

Ниже приведен пример ($n = 50$) данных о высоте деревьев на разных широтах. Каждая точка соответствует одному дереву.

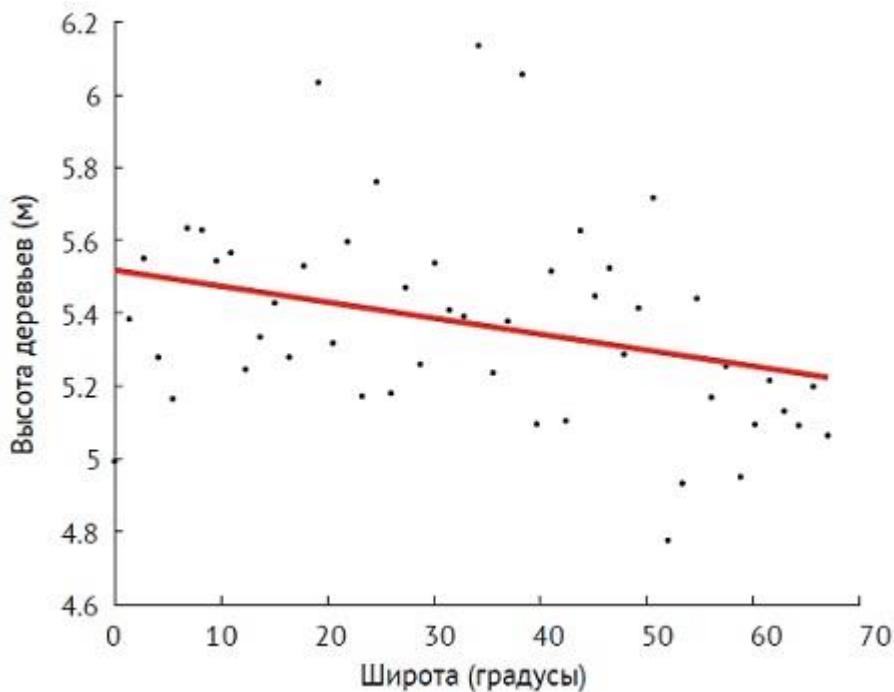


Рис. 20. Зависимость высоты деревьев от географической широты для выборки из 50 деревьев. Коэффициент корреляции $r = -0.312$. Красная линия – наилучшая эмпирическая прямая

Проверим гипотезу

$$H_0: \rho = 0,$$

где ρ – корреляция в генеральной совокупности.

Если нулевая гипотеза верна, то стандартное отклонение выборочного распределения выборочной корреляции равно:

$$(18) s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

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

$$(19) t = \frac{r - 0}{s_r}$$

с числом степеней свободы $df = n - 2$. В этом случае статистические программы покажут:

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.

Интерпретация корреляции

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.

Корреляция измеряет только линейную связь, поэтому незначимая корреляция еще не означает, что между x и y нет никакой причинно-следственной связи. 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 , поэтому знание одной лишь корреляции несет только частичную информацию о наборе данных.

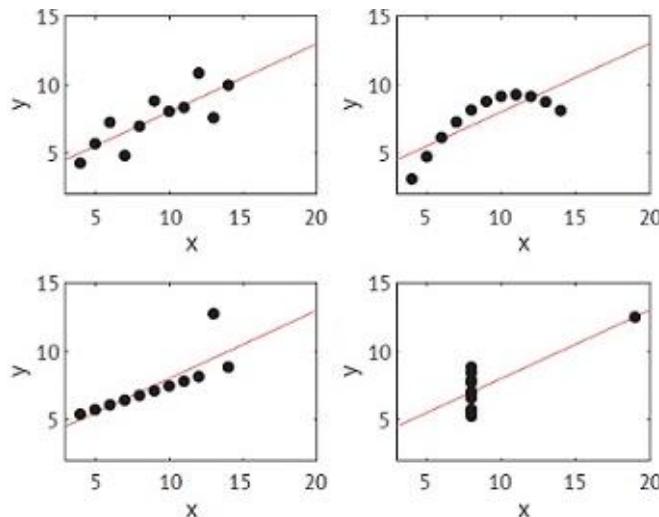


Рис. 22. [Квартет Энскомба](#). Для всех наборов данных значение корреляции одно и то же $r = 0,816$

Корреляция очень чувствительна к выбросам, и добавление или удаление всего одной точки может изменить коэффициент корреляции набора данных.

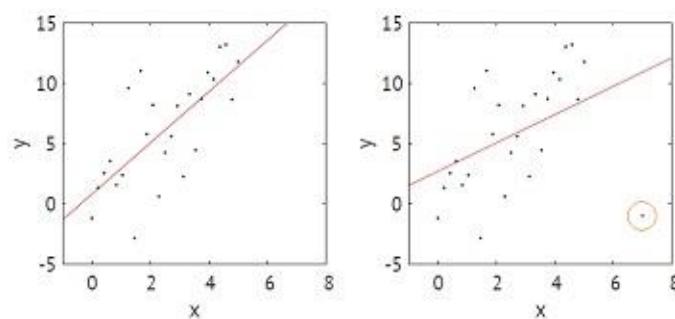


Рис. 23. Влияние выбросов на корреляцию. Слева: оригинальные данные с $r = 0,71$. Справа: добавление единственного выброса привело к уменьшению $r = 0,44$

	Малый	Средний	Большой
Размер эффекта	0.1	0.3	0.5

Рис. 24. Рекомендации Коэна по размеру эффекта для r

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

Часть III. Метаанализ и кризис науки

Глава 9. Метаанализ

Путем комбинирования данных из различных экспериментов можно получить новые знания.

Насколько вероятно, что все четыре эксперимента с малым эффектом и малым размером выборки приведут к значимым результатам? Крайне маловероятно. Отсюда вытекает простое следствие: если эксперименты всегда дают значимые результаты, то данные слишком хороши, чтобы быть правдой. Распространенная, но неправильно понятая научная практика ведет к «слишком хорошим, чтобы быть правдой» данным. Такая практика вздувает частоту ошибок I-го рода. Это привело к серьезному кризису науки, затронувшему многие области, где статистика играет ключевую роль.

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 Коэна, а g Хеджеса:

$$(20) g = \left(1 - \frac{3}{4(2n - 2) - 1}\right) d$$

Предположим, что похожие эксперименты выполняются несколько раз. Для проведения метаанализа необходимо взвесить результаты отдельных экспериментов. Вес (w) зависит от размера выборки через дисперсию g Хеджеса – v_g :

$$(21) w = \frac{1}{v_g}$$

Дисперсия g Хеджеса для случая равенства размеров сравниваемых в каждом эксперименте выборок $n_1 = n_2 = n$

$$(22) v_g = \left(1 - \frac{3}{4(2n - 2) - 1}\right)^2 \cdot v_d$$

где v_d – дисперсия d Коэна

$$(23) v_d = \frac{2n}{n^2} + \frac{d^2}{4n}$$

	A	B	C	D	E	F	G	H
1	n	t	d	g	v_d	v_g	w	w^*g
2	36	3,01	0,709	0,702	0,059	0,058	17,305	12,15
3	36	2,08	0,490	0,485	0,057	0,056	17,857	8,66
4	36	2,54	0,599	0,592	0,058	0,057	17,605	10,43
5	46	3,08	0,642	0,637	0,046	0,045	22,243	14,17
6	46	3,49	0,728	0,722	0,046	0,046	21,937	15,83
7								
8				g^*	0,632			

Рис. 25. Шаги метаанализа

Объединенный размер эффекта вычисляется путем сложения взвешенных размеров эффектов и деления результата на сумму весов:

$$(22) g^* = \frac{\sum_{i=1}^k w_i g_i}{\sum_{i=1}^k w_i}$$

где k – число экспериментов в метаанализе.

Метааналитический размер эффекта $g^* = 0,632$ – лучшая оценка, основанная на результатах пяти экспериментов.

Глава 10. Воспроизведимость

Во всех науках воспроизводимость считается «золотым стандартом». К сожалению, во многих научных дисциплинах с воспроизводимостью дело обстоит неважно. Группа психологов [Open Science Collaboration](#) 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. 26 представляет пару p -значений: из оригинального и повторного исследования. Штриховая вертикальная прямая обозначает порог 0,05 в оригинальных исследованиях. 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 оригинальным 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.

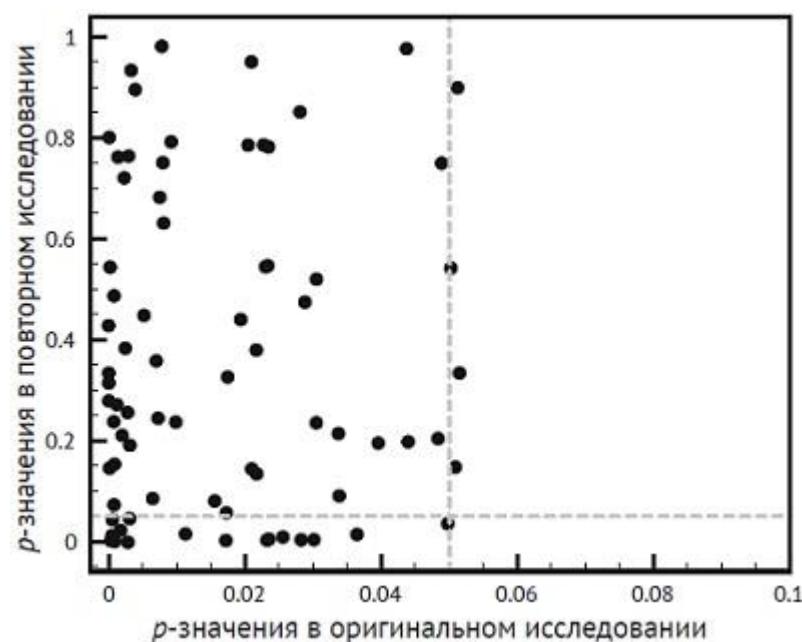


Рис. 26. Каждая точка соответствует паре p -значений из оригинального исследования и его воспроизведения. Обратите внимание на различный масштаб по осям x и y . Хорошим считался бы результат, при котором p -значения в повторном исследовании были меньше 0,05, т. е. когда все черные точки расположены ниже штриховой горизонтальной прямой

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. 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.

Избыточный успех может возникать из-за того, что публикуются только эксперименты, демонстрирующих значимые результаты. Второй источник статистического смещения – изменение размера выборки в процессе эксперимента.

Например, исследователь собирает данные из двух генеральных совокупностей, и в результате в каждой выборке оказывается $n_1 = n_2 = 10$ элементов. Исследователь вычисляет t -критерий и находит, что $p = 0,08$. Это выше порога статистической значимости, и... исследователь может взять еще десять

элементов, чтобы в каждой выборке было $n_1 = n_2 = 20$. Предположим, что t-критерий дает $p = 0,04$. В таком виде и публикуются итоги исследования.

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.

Если каждый из многих похожих экспериментов с малыми размерами эффекта и выборки дает значимый результат, то данные слишком хороши, чтобы быть правдой. Количество значимых результатов в экспериментах должно быть пропорционально их мощности. Статистическое смещение публикации и необязательная остановка могут сильно увеличить частоту ошибок I-го рода, т.е. вероятность подтверждения альтернативной гипотезы, когда всё еще верна нулевая.

Глава 11. Величина избыточного успеха

Выявить проблемы в данных позволяет тест избыточного успеха. Его идея основана на том, что неудачи обязательно должны быть. Даже если эффект реально существует, иногда должны встречаться выборки, в которых он не проявляется. Смоделируем три серии экспериментов.

Каждый набор данных анализировался с помощью двустороннего t-критерия. Во всех экспериментах размеры выборок равны ($n_1 = n_2 = n$). В одной серии (корректная проверка) размер эффекта в генеральной совокупности был задан 0,8, а n случайно выбирались из диапазона 10–30. Все пять экспериментов дали значимые результаты, которые были полностью опубликованы.

В следующей серии размер эффекта в генеральной совокупности равен 0 (эффект отсутствует). Выборка генерировалась с применением необязательной остановки: вначале $n = 10$, а затем увеличивалось с шагом 1 до достижения максимума 30. Всего было выполнено 20 экспериментов, пять из которых дали значимый результат. Только эти пять значимых экспериментов и были опубликованы.

В третьей серии размер эффекта был задан равным 0,1. n случайно выбиралось из диапазона 10–30. Всего было выполнено 100 экспериментов, из которых пять дали значимый результат. Только они и были опубликованы.

Прежде чем читать далее, изучите статистику и примите решение, какая из выборок корректна.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Серия А				Серия В				Серия С			
2	n	t	p	g	n	t	p	g	n	t	p	g
3	10	2,620	0,028	1,122	23	2,884	0,009	0,836	22	2,181	0,041	0,646
4	13	2,223	0,046	0,844	30	5,599	0,000	1,427	28	2,511	0,018	0,662
5	28	2,056	0,050	0,542	15	2,711	0,017	0,963	11	2,298	0,044	0,943
6	22	2,267	0,034	0,671	18	4,979	0,000	1,623	12	2,223	0,048	0,876
7	24	2,253	0,034	0,640	26	2,527	0,018	0,690	19	3,992	0,001	1,268

Рис. 27. Сводная статистика для трех серий из пяти экспериментов. В одной серии была использована необязательная остановка, в другой – отбор экспериментов, а третья выполнена корректна. В какой серии нет подвоха?

Нашли корректную серию? Если задание показалось вам трудным или вы не уверены в правильности ответа, не расстраивайтесь – это сложный вопрос. Можно применить метаанализ с объединением размеров эффектов по 5 экспериментам в каждой серии. Для серии А $g^* = 0,69$, для серии В $g^* = 1,07$, а для серии С $g^* = 0,83$. (Я провел расчеты в приложенном Excel-файле. – Прим. Багузина.) Но и эта информация не помогает идентифицировать корректную серию экспериментов.

Здесь полезен тест избыточного успеха. Детали вычислений приведены в Excel-файле на листе "Мощность". Мощность для каждого эксперимента можно вычислить с помощью программы G*Power (см. рис. 18) или воспользовавшись формулами в Excel. Если отдельно по каждой серии перемножить мощности пяти экспериментов, то получится, вероятность, с которой все пять

экспериментов дают значимые результаты. Для серии А $p = 0,054$, для серии В $p = 0,376$, для серии С $p = 0,094$. Очень похоже, что корректной является серия В.

Тест избыточного успеха – это формальный анализ, но существуют эвристические правила, которые позволяют оценить корректность серии экспериментов. Можно посмотреть на связь между размером выборки и размером эффекта. В корректной серии эти величины не коррелируют (больший размер выборки ведет к более точным оценкам размера эффекта, но на сам размер эффекта не влияет). Для серии В на рис. 27 корреляция между размером выборки и размером эффекта $r = -0,05$, что отражает случайность размера эффекта в разных экспериментах. Напротив, для серий А и С $r = -0,95$ и $-0,44$ 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).

Еще один признак проблематичного набора данных – когда многие p -значения близки к порогу статистической значимости, но 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 В на рис. 27 почти все p -значения очень малы, а в двух других сериях 8 из десяти p -значений от 0,02 до 0,05. Такое распределение p -значений должно насторожить.

[Насколько широко распространены эти проблемы?](#)

Чтобы исследовать, насколько распространены статистически недостоверные эксперименты, мы обратились к онлайновой системе поиска по журналу *Science* и нашла 133 статьи по психологии или образованию, опубликованных в период между 2005 и 2012 годом. Мы применили тест избыточного успеха к каждой из 18 статей, в которых было описано четыре и более эксперимента и содержалось достаточно информации для оценивания вероятности успеха. В 15 из 18 (83 %) статей в *Science* результаты оказались слишком хорошими, чтобы быть правдой (т. е. вероятность успеха в них была менее 0,1).

Похоже 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.

[Глава 12. Предлагаемые улучшения и нерешенные проблемы](#)

Критики проверки гипотез иногда заявляют, что p -значения, используемые в этой методологии, зашумлены или ни о чем не говорят. Однако p -значение основано на информации, присутствующей в наборе данных. p -значение оценивает отношение сигнала к шуму, как 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.

В некотором смысле долгосрочной целью науки является устранение статистики путем отыскания «правильных» факторов и уменьшения изменчивости. 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 достижению 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.