Дата последней сборки документа:
7 февраля 2024 г. в 23:51

Доверительное A/B-тестирование

Содержание

1. Вводные замечания

Контролируемые эксперименты называют A/B-тестами, A/B/n-тестами, полевыми экспериментами, *рандомизированными контролируемыми экспериментами*, сплит-тестами, тестами с корзиной и пробными полетами.

Каждый год крупные компании проводят до десятков тысяч экспериментов с участием миллионов пользователей и тестируют все подряд, включая изменения пользовательского интерфейса, алгоритма релевантности (поиск, реклама, персонализация, рекомендации и т.д.), задержку/-быстродействие, системы управления контентом, системы поддержки клиентов и многое другое. Эксперименты проводят по нескольким каналам: веб-сайтам, приложениям для компьютеров, мобильным приложениям и электронной почте.

Общий критерий оценки (OEC) – количественный показатель цели эксперимента. Например, ваш OEC может быть выражен в днях на пользователя, указывая количество дней во время эксперимента, в течение которых пользователи были активны (то есть они посетили сайт и

предприняли какие-либо действия). OEC должен быть измеримым в краткосрочной перспективе (продолжительность эксперимента), но при это иметь причинное влияние на долгосрочные стратегические цели.

Параметр – контролируемая экчспериментальная переменная, которая, как ожидается, влияет на OEC или другие интересующие показатели. В простых A/B-тестах обычно используют один параметр с двуями значениями. В онлайн-тестировании чаще используют конструкции с одной переменной и несколькими значениями (например, A/B/C/D). Многовариантные тесты, также называемые мультивариантыми, оценивают вместо несколько параметров, таких как цвет и размер шрифта, что позволяет экспериментатрам находить глобальный оптимум при взаимодействии параметров.

Вы должны сопоставлять объекты с вариантами постоянным и и независимым образом (то есть если объектом рандомизации является пользователь, он должен постоянно видеть один и тот же опыт, а назначение пользователя определенному варианту не должно ничего говорить вам о назначении другого пользователя другому варианту). Некоторые экспериментальные проекты выбирают рандомизацию по страницам, сеансам или пользовательским дням (то есть эксперимент остается неизменным для пользователя в течение каждого 24-часового окна, определенного сервером).

Контролируемые онлайн-эксперименты:

- являются лучшим научным способом выявить *причино-следственную связь* с высокой степенью достоверности,
- позволяют обнаруживать небольшиме изменения, которые труднее обнаруживать другими методами, например изменение во времени (чувствительности),
- позволяют обнаруживать неожиданные изменения.

1.1. Необходимые ингредиенты для проведения эффективных контролируемых экспериментов

Для проведения контролируемых экспериментов нужны следующие компоненты и условия [?, стр. 35]:

- экспериментальные объекты (например, пользователи), которые могут быть отнесены к разным вариантам без перекрестного взаимного влияния (или с небольшим влиянием); например, пользователи в тестовой группе не влияют на пользователей в контрольной группе,
- достаточное количество экспериментальных объектов (например, пользователей). Чтобы контролируемые эксперименты были полезными, мы рекомендуем использовать тысячи экспериментальных объектов: чем больше число, тем тоньше эффекты, которые можно обнаружить,
- ключевые показатели, в идеале OEC, сформулированы и могут быть измерены на практике. Если реальные цели слишком сложно измерить, важно договориться о *суррогатных целях*. Вы должны иметь возможность собирать надежные данные, в идеале – дешево и масштабно.
- простота внесения изменений.

Показатель «прибыль» – не лучший вариант OEC, поскольку сиюминутные действия (например, повышение цен) могут увеличить краткосрочную прибыль, но и навередить ей в долгосрочной перспективе. И наоборот, *длительность жизненного цикла клиента* (customer lifetime value, CLV) – это стратегически мощный OEC. Трудно переоценить важность разработки хорошего OEC, который подходит именно вашей организации.

Не рекомендуется использовать в качестве показателя саму сумму дохода, поскольку она зависит от количества пользователей в каждом варианте. Даже если вы стараетесь поровну распределить пользователей между вариантами, фактическое количество пользователей может измениться случайно. Мы рекомендуем нормализовать ключевые показатели по фактическим размерам выборки, чтобы получить *доход на пользователя*.

Следующий важный вопрос – решить, каких пользователей следует учитывать в знаменателе при вычислении дохода на пользователя:

- всех пользователей, посетивших сайт: это допустимый вариант, однако, он дает размытую выборку, потому что в нее входят пользователи, которые никогда не начинали оформление заказа в варианте, где было внесено изменение; исключение этих пользователей приведет к более чувствительному A/B-тестированию,
- только пользователей, завершивших процесс заказа: этот выбор неверен, поскольку предполагает, что изменение повлияет на сумму покупки, а не на процент пользователей, совершивших покупку; если покупает больше пользователей, доход на пользователя может упасть, даже если совокупный доход увеличится,
- только пользователей, которые начинают оформление заказа: это лучший выбор с учетом того, где находится изменение в воронке; мы включаем всех потенциально затронутых пользователей, но исключаем незатронутых (пользователей, которые никогда не начинали оформление заказа), искажающих наши результаты.

Мы количественно проверяем вероятность совпадения между тестовой и контрольной выборкой исходя из *нулевой гипотезы* о том, что H_0 : «средние значения совпадают».

P-значение – вероятность наблюдения такой или более экстремальной разницы, если предположить, что нулевая гипотеза верна. Мы отвергаем нулевоую гипотезу и делаем вывод, что наш эксперимент имеет эффект (или результат статистически значим), если p-значение достаточно мало. Общепринятый научный стандарт – использовать значение p меньше 0.05. Это означает, что, если изменение действительно не дает эффекта, мы можем правильно сделать вывод об отсутсвии эффекта в 95 случаях из 100 [2, стр. 57].

Другой способ проверить, является ли разница статистически значимой, – это проверить, не перекрывает ли *доверительный интервал* нулевое значение. Доверительный интервал 95% – это диапазон, который покрывает истинную разницу в 95% случаев, и для достаточно большой выборки он обычно сосредоточен вокруг наблюдаемой дельты между тестом и контролем с расширением 1.96 стандартных ошибок с каждой стороны.

То есть, если p-значение меньше уровня значимости α, мы заявляем, что разница *статистически значима* $\text{p-value} < \alpha \to \cancel{H_0}$. Эквивалентное представления использования *95% доверительного интервала* $[\Delta - 1.96\sigma, \Delta + 1.96\sigma]$ для оценки статистической значимости (Δ – наблюдаемая разница). Если ноль лежит за пределами доверительного интервала, мы заявляем о *статистической значимости* $0 \notin [\Delta - 1.96\sigma, \Delta + 1.96\sigma] \to \cancel{H_0}$ [2, стр. 58].

Статистическая мощность (statistical power) – это вероятность обнаружения значимого различия между вариантами, когда оно действительно есть. С практической точки зрения вам нужно, что ваш эксперимент обладал достаточной статистической мощностью, чтобы с высокой вероятностью сделать вывод о том, привел ли ваш эксперимент к более значительным изменениям, чем вы предполагали. Обычно чем больше выборка, тем больше мощность. Обычной практикой является планирование экспериментов для мощности 80%–90% [2, стр. 57].

Хотя «статистическая значимость» измеряет, насколько вероятно, что наблюдаемый вами или более выраженный результат мог возникнуть случайно при условии, что он равен нулю, не все *статистически значимые* результаты являются значимыми на практике. Насколько большая разница, в данном случае изменение дохода на пользователя. действительно имеет значение для нас с точки зрения бизнеса? Другими словами, какое изменение *практически значимо*?

Определение границы значимости важно для понимания того, стоит ли разница затрат на внесение изменений.

1.2. Разработка эксперимента

У нас есть гипотеза, граница практической значимости, и мы определили метрку. Теперь, чтобы завершить разработку эксперимента, мы должны ответить на следующие вопросы:

1. Какова у нас единица рандомизации?
2. Какую популяцию единиц рандомизации мы хотим охватить?
3. Насколько большим (по охвату) должен быть наш эксперимент?
4. Как долго мы проводим эксперимент?

Предположим, что наша единица рандомизации – это пользователь. Ориентация на конкретную популяцию означает, что вы хотите провести эксперимент только для пользователей с определенной характеристикой. Например, вы тестируете новый текст, представленный лишь на нескольких языках; в этом случае вы можете ориентироваться только на пользователей, у которых в языковой настройке интерфейса выбран один из этих языков.

Размер эксперимента (для нас количество пользователей) напрямую влияет на точность результатов. Если вы хотите обнаружить небольшое изменение или быть более уверенным в своих выводах, проводите более масштабный эксперимент с большим количеством пользователей. Для повышения точности можно использовать и другие изменения параметров эксперимента:

- если мы будем использовать индикатор покупки (то есть сделал ли пользователь покупку, да/нет, без учета суммы покупки) вместо пользования дохода на пользователя в качестве нашего OEC, стандартная ошибка будет меньше, а это означает, что нам не нужно подвергать эксперименту как можно большее количество пользователей, чтобы достичь той же чувствительности,
- если мы увеличим наш *практический порог значимости*, сказав, что больше не заботимся об обнаружении 1% изменения, а только о более значительных изменениях, мы могли бы уменьшить размер выборки, потому что *более крупные изменения легче обнаружить*,
- если мы хотим использовать более низкий порог p-значения, например 0,01, чтобы быть более уверенными в наличии эффекта до того, как мы отклоним нулевую гипотезу, нам необходимо *увеличить размер выборки*.

Другой большой вопрос – как долго проводить эксперименты. Есть и другие факторы, которые следцет учитывать:

- рост числа пользователей,
- эффект дня неделя; даже один и тот же пользователь может вести себя по-разному в разным дни недели. Рекомендуется проводить эксперименты как минимум в течение одной недели,
- сезонность; могут быть и другие моменты, когда пользователи ведут себя иначе, например в праздники,

- эффекты первенства и новизны; существуют эксперименты, которые, как правило, дают больший или меньший начальный эффект, для стабилизации которого требуется время. Например, пользователи могут попробовать нажать новую яркую кнопку и обнаружить, что она бесполезна, поэтому количество нажатий на кнопку со временем будет уменьшаться. С другой стороны, функции, требующие привыкания, требуют времени для содания группы приверженцев.

В целом чрезмерная статистическая мощность эксперимента – это хорошо и даже рекомендуется, поскольку иногда нам нужно исследовать сегмент пользователей (например, географический регион или платформу) и убедиться, что эксперимент имеет достаточную мощность для обнаружения изменений нескольких ключевых показателей.

В эксперимент могут вкраться ошибки, нарушающие достоверность результатов. Чтобы выявить их, мы посмотрим на *инвариантные* или ограничительные показатели. *Эти показатели должны быть одинаковыми в контрольной и тестовой группе*. Если они меняются, любые измеренные различия, скорее всего, являются результатом других изменений, которые мы внесли, а не тестируемой функции.

Есть два типа инвариантных показателей:

- Доверительные ограничительные показатели, такие как ожидание того, что контрольные и тестовые выборки будут иметь размер в соответсвии с конфигурацией эксперимента или что они имеют одинаковую частоту попаданий в кэш.
- Организационные ограничительные показатели, такие как задержка, которые важны для организации и, как ожидается, будут инвариантном для многих экспериментов. В эксперименте с оформлением заказа было бы очень странно, если бы задержка обработки заказа изменилась от добавления фиктивного поля.

Если эти проверки достоверности не пройдены, вероятно, проблема кроется в основном плане эксперимента, инфраструктуре или обработке данных.

1.3. Неправильная интерпретация статистических результатов

1.3.1. Нехватка статистической мощности

В нашей методике проверки значимости нулевой гипотезы мы обычно предполагаем, что нет разницы в значениях показателей между контрольной и тестовой группой (нулевая гипотеза), и отклоняем гипотезу, если данные представляют убедительные доказательства против нее.

Распространненой ошибкой является вывод об отсутствии эффекта от изменения лишь на том основании, что показатель не является статистически значимым. Вполне возможно, что эксперименту *не хватает статистической мощности* для определения размера эффекта, который мы наблюдаем, то есть для успешного теста просто не хватило пользователей [2, стр. 67].

Необходимо определить порог практической значимости в вашем контексте, и убедится, что у вас достаточно мощности эксперимента, чтобы обнаружить изменение показателя на такую или меньшую величину.

1.3.2. Неправильная интерпретация p-значений

p-значение – это вероятность получения результата, равного или более экстремального, чем наблюдаемый, при условии, что нулевая гипотеза верна [2, стр. 68]. Условие нулевой гипотезы имеет решающее значение.

1.3.3. Множественные проверки гипотез

Когда существует несколько тестов и мы выбираем наименьшее значение p, наши оценки значения p и величины эффекта, скорее всего, будут смещены. Это происходит в следующих случаях:

- отслеживание нескольких показателей,
- отслеживание p-значений во времени,
- анализ сегментов популяции (например, страна, тип браузера, новый/постоянный клиент),
- рассмотрений нескольких итераций эксперимента. Например, если эксперимент действительно ничего не дает (A/A), его выполнение 20 раз подряд может случайно привести к p-значению меньше 0.05.

Коэффициент ложного обнаружения (false discovery rate) – ключевая концепция для работы с несколькими тестами.

1.4. Доверительные интервалы

Доверительный интервал (confidence interval), грубо говоря, количественно определяет *степень неопределенности* эффекта от изменения (границы доверительного интервала являются случайными величинами). Уровень доверия показывает, как часто доверительный интервал должен содержать истинный эффект от воздействия.

Между p-значениями и доверительными интервалами существует определенное родство. Для нулевой гипотезы об отсутствии разницы, обычно используемой в контролируемых экспериментах, 95%-ный доверительный интервал эффекта воздействия, который не пересекает ноль, означает, что $p < 0.05$ [2, стр. 70].

Распространенной ошибкой является рассмотрение доверительных интервалов отдельно для контрольной и тестовой группы и предположение, что, если они перекрываются, изменение не дает статистически значимого эффекта. Это неверно. Доверительные интервалы могут перекрываться до 29%, но при этом разница будет *статистически значимой*[2, стр. 70].

Однако верно и обратное: если 95%-ные доверительные интервалы не перекрываются, то эффект от воздействия статистически значим с $p < 0.05$ [2, стр. 70].

1.5. Угрозы внутреней достоверности

1.5.1. Несоотетствие коэффициента выборки

Если соотношение пользователей между вариантами значительно отличается от задуманного значения, эксперимент страдает от *несоответствия коэффициента выборки* (sample ratio mismatch, SRM).

При больших числах коэффициент меньше 0.99 или больше 1.01 для плана эксперимента, который требует 1.0, скорее всего, указывает на серьезную проблему.Как было сказано ранее, p-значение – это вероятность получения результата, равного или более экстремального, чем наблюдаемый, при условии, что гипотеза о нуле верна. Если план эксперимента предусматривал равные распределения для обоих вариантов, то по плану вы должны получить соотношение близкое к 1.0, то есть гипотеза о нуле должна быть верной. Таким образом, p-значение представляет собой вероятность того, что наблюдаемое соотношение согласуется с планом нашего эксперимента.

Проверка SRM критически важна. Даже небольшой дисбаланс может привести к обратному эффекту от тестового воздействия. Сдвиги SRM обычно возникают из-за исключения пользователей (как правило, экспериментальных единиц), которые либо очень хороши, например интенсивные пользователи, либо очень плохи, как пользователи, у которых нет отмеченных кликов.

Это говорит о том, что, даже если разница в численности популяции кажется небольшой, она может существенно исказить результаты.

1.6. Угрозы внешней достоверности

Внешняя достоверность (external validity) отражает степень, в которой результаты контролируемого эксперимента могут быть обобщены по направлениям, таким как различные группы населения (например, другие страны, другие веб-сайты) и с течением времени (например, будет ли увеличение дохода на 2% удерживаться в течение длительного времени, или доход постепенно уменьшится).

Обобщения во времени даются еще сложнее. Иногда после эксперимента приходится ждать несколько месяцев, пока утихнут долгосрочные эффекты.

1.7. Парадокс Симпсона

Если мы проводим эксперимент с накоплением, то есть имеем два или более периода с *разными процентными долями* распределения вариантов, объединение результатов может привести к смещенной неверной оценке эффектов воздействия, то есть тест может быть лучше, чем контроль как в первой, так и во второй фазе, но в целом хуже, когда два периода объединены. Это явление называется *парадокс Симпсона* [2, стр. 82].

Нужно быть осторожными при агрегировании данных, собранных в *разных процентных отношениях*.

Если требуется максимальная мощность, эксперимент будет проводиться при процентном соотношении 50/50 и включать всех пользователей.

OEC должен быть измеримым в краткосрочной перспективе (продолжительность эксперимента), но прир этом оставаться причинно-следственным фактором для достижения долгосрочных стратегических целей.

Каждый из сотен или даже тысяч экспериментов, проведенных в прошлом, представляет собой страницу в журнале с ценными и обширными данными о каждом изменении (реализованом или нет). Этот цифровой журнал мы называем *институциональной памятью* (institutional memory). Настоятельно рекомендуется собирать общую информацию по каждому эксперименту, например кто является инициатором; когда начался эксперимент; как долго он работал; описания и снимки экрана, если изменение было визуальным. Наконец, вы должны завершить гипотезу, на которой основан эксперимент: какое решение было принято и почему.

1.8. Планы для наблюдательных исследований причино-следственных связей

Планы:

- прерывистый временной ряд,
- эксперименты с чередованием,
- метод разрывной регрессии,
- инструментальные переменные и естественные эксперименты,

- отбор подобного по сколнности,
- дифференциальная разница.

Прерывистый временной ряд (Interrupted time series) – это квазиэкспериментальный план, в котором вы можете контролировать изменения в своей системе, но не можете рандомизировать воздействие, чтобы иметь надлежащую контрольную и тестовую популяцию. Вместо этого вы используете одну и ту же популяцию для контроля и тестирования и меняете то, что она испытывает с течением времени.

В частности, план использует несколько измерений во времени до воздействия для создания модели, которая может предоставить оценку интересующего показателя после воздействия – то есть контрфактуальную. После воздействия также проводится несколько измерений, и эффект воздействия оценивается как средняя разница между фактическими значениями интересующего показателя и значениями, предсказанными моделью.

Одним из дополнений к простому плану ITS является применение воздействия, а затем его отмена; при необходимости эту процедуру повторяют несколько раз.

Одна из распространенных проблем, связанных с исследованиями причино-следственных связей, заключается в том, чтобы гарантировать, что вы не приписываете какой-либо эффект изменению, хотя на самом деле имеется некоторый совместный эффект.

План *экспериментов с чередованием* – это распространенный план, используемый для оценки изменений алгоритма ранжирования, например в поисковых системах или при поиске на вебсайте. Хотя это мощный экспериментальный план, его применимость ограничена, поскольку результаты дложны быть однородными. Если, как это часто бывает, первый результат занимает больше места или влияет на другие области страницы, возникают сложности.

Метод разрывной регрессии (regression discountinuty design, RDD) – это план эксперимента, который можно использовать всякий раз, когда есть четкий порог, который идентифицирует исследуемую популяцию. Основываясь на этом пороговом значении, мы можем уменьшить системетическую ошибку отбора, определяя популяцию, которая чуть ниже порога, как контрольную, а популяцию, которая выше порога, как тестовую.

Если стипендия предоставляется учащимся, набравшим 80% от максимального балла, то предполагается, что тестовая группа, получившая оценки чуть выше 80%, аналогична контрольной группе, получившей оценки чуть ниже 80%.

Пример использования RDD – оценка влияния употребления алкоголя на смертность: американцы старше 21 года могуть пить алкгогольные напитки легально, поэтому мы можем рассмотреть количество сметрей по дням рождения. Скачок в возрасте 21 года, похоже, не является обычным следствием празднования дня рождения. Если бы этот всплеск отражал просто вечеринки по случаю дня рождения, то мы наблюдали бы увеличение числа сметрей и после 20-го и 22-го дня рождения, но этого не происходит [2, стр. 179].

Ключевой проблемой снова является наложение факторов. В RDD поведение, показателей может быть искажено другими факторами, которые связыны с тем же порогом. Например, исследование влияния алкоголя, в котором в качество порогового значения выбирается возраст, равный 21 году, может быть загрязнено тем фактом, что это также порог для легального участия в азартных играх.

Инструментальные переменные (instumental variables, IV) – это метод, который пытается аппроксимировать случайное назначение. В частности, цель состоит в том, чтобы найти инструмент, который позволяет нам аппроксимировать случайное распределение.

Например, при анализе разницы в заработках веетеранов и неветеранов выборочный призыв на войну во Вьетнаме аппроксимирует случайный призыв людей в армию.

Отбор подобного по склонности состоит в построении сопоставимых контрольных и подопытных групп, часто путем сегментации пользователей на основании общих специфических свойств или склонностей к чему-либо, – это что-то вроде стратифицированной выборки.

Идея состоит в том, чтобы гарантировать, что различие между контрольной и подопытной популяции не связано с изменением состава популяции. Например, если мы изучаем экзогенное изменение воздействия перехода пользователей с Windows на IOS, то хотим убедиться, что мы не измеряем демографическую разницу в населении.

Мы можем продолжить этот подход, перейдя к отбору подобного по склонности (proppensity score matching, PSM), которое вместо сопоставления единиц на ковариатах сопоставляет одно число: сформированную оценку склонности. Этот подход использовался в онлайн-пространстве, например для оценки воздействия рекламных компаний в интернете. Оценка предрасположенности работает только в условиях «сильного игнорирования». Для всех этих методов основной проблемой является наложение факторов.

Метод дифференциальной разницы (difference in differences, DD) учитывает разницу в различиях контрольной и подопытной группе. Иными словами, группы могут различаться при отсутсвии воздействия, но двигаться параллельно. Этот метод обычно используется в географических экспериментах.

1.9. Ловушки причинно-следственных связей

Хотя наблюдательные исследования причино-следственных связей иногда являются наилучшими вариантом, они скрывают много подводных камней. Основная ошибка, независимо от метода, при проведении наблюдательных исследований причинно-следственных связей – это *непредвиденные заблуждения*, которые могут повлиять на измеряемый эффект или привести к попытке объяснить причинностью заурядную смену интересов. Из-за этих заблуждений наблюдательные исследования причинно-следственных связей требуют большой осторожности.

Один из распространенных типов заблуждений – это нераспознанная *общая причина*. Например, у людей размер ладони сильно коррелирует с продолжительностью жизни: в среднем чем меньше ваша ладонь, тем дольше вы проживете. Однако общей причиной меньших размеров ладоней и большей продолжительности жизни является пол: женщины имеют меньшие ладони и в среднем живут дольше. То есть между размером ладони и продолжительностью жизни не существует зависимости; пол является общей причиной, объясняющей и то и другое.

Другой пример: для многих сервисов, в том числе Microsoft Office 365, чем больше пользователей сталкивается с ошибками, тем меньше они покидают сервис. Но не пытайтесь увеличить число ошибок, ожидая уменьшения оттока пользователей, так как эта корреляция обусловлена общей причиной: интенсивностью использования. Самые активные пользователи чаще сталкиваются с ошибками, но и уходят реже. Чтобы оценить, действительно ли новая функция снижает отток, запустите управляемый эксперимент (и проанализируйте новых и активных пользователей по отдельности).

Еще одна ловушка, о которой следует помнить, – это *ложные*, или *обманчивые, корреляции*. Обманчивая корреляция может быть вызвана сильными выборосами. Например маркетинговая компания может заявить, что их энергетический напиток сильно коррелирует со спотривными

результатами и подразумевает причинно-следственную связь: дескать, выпейте наш напиток, и выши результаты улучшатся.

Заявление о причино-следственной связи на основе *неконтролируемых* (наблюдательных) экспериментов требует множества предположений, которые невозможно проверить и которые легко нарушаются [2, стр. 185].

Более активные пользователи просто с большей вероятностью будут выполнять более широкий круг действий. Как правило, важно использовать активность как фактор.

2. Статистика контролируемых онлайн-экспериментов

2.1. Двухбыорочный t-тест

Двухвыборочный t-тест является наиболее распространенным критерием статистической значимости для определения того, является ли разница, которую мы видим между тестовой и контрольной группой, реальной или просто шумом. Двухвыборочный t-тест оценивает размер разницы между двумя средними значениями относительно дисперсии. Значимость разницы представлена p-значением. Чем ниже значение p, тем сильнее доказательство того, что тестовая группа отличается от контрольной.

Чтобы применить двухвыборочный t-тест к интересующему показателю Y (например, числу запросов на пользователя), предположим, что наблюдаемые значения показателя для пользователей в контрольной и тестовой группе являются независимыми реализациями случайных величин Y^t и Y^c. Гипотеза нулевого значения состоит в том, что Y^t и Y^c имеют одинаковое среднее значение; альтернативная гипотеза состоит в том, что это не так.

$$H_0 : mean(Y^t) = mean(Y^c),$$
$$H_1 : mean(Y^t) \neq mean(Y^c)$$

Интуитивно понятно, что чем больше T, тем меньше вероятность того, что средние значения совпадают.

2.2. p-значение и доверительный интервал

Теперь, когда есть t-статистика T, вы можете вычислить p-значение, которое представляет собой вероятность того, что T будет, по крайней мере, иметь такое значение, если действительно нет разницы между группами.

По соглашению, любое различие с p-значением меньше 0.05 считается «статистически значимым», хотя продолжаются дискуссии о необходимости более низких p-значений по умолчанию. Значение p меньше 0.01 считается очень значимым.

Другой способ проверить, является ли дельта статистически значимой, – это проверить, не перекрывается ли доверительный интервал с нулем.

Доверительный интервал 95% – это диапазон, который покрывает истинную разницу в 95% случаев и соответствует p-значению 0.05; дельта является статистически значимой на уровне значимости 0.05, если 95%-ный доверительный интервал не содержит нуля или если $p < 0.05$. В большинстве случае доверительный интервал для дельты центрируется вокруг наблюдаемой дельты с расширением примерно на два стандартных отклонения с каждой стороны. Это верно

для любой статистики, которая (приблизительно) соответствует нормальному распределению [2, стр. 231].

В большинстве случаев мы вычисляем p-значение с предположением, что t-статистика T следует нормальному распределению и, согласно, гипотезе о нуле, распределение имеет среднее 0 и дисперсию 1. Значение p – это просто площадь под нормальной кривой. В большинстве онлайн-экспериментов размеры выборок как для контроля, так и для тестирования исчисляются как минимум тысячами. Хотя выборочное распределение Y не следует нормальному распределению, среднее $\bar{Y}$ обычно следует ему в силу центральной предельной теоремы.

Одно практическое правило для минимального количества образцов, необходимых для среднего $\bar{Y}$ для нормального распределения, составляет $335 \cdot s^2$, где s – коэффициент ассиметрии выборочного распределения Y [2, стр. 232].

Это практическое правило дает хорошее представление о том, когда $|s| > 1$, но не дает полезной нижней границы, если распределение симметрично или имеет небольшую асимметрию. С другой стороны верно то, что при меньшей асимметрии требуется меньше выборок.

Крайне важно не только правильно оценить дисперсию, но и понять, как добиться уменьшения дисперсии, чтобы повысить чувствительность тестирования статистических гипотез.

При проведении контролируемого эксперимента мы стремимся обнаружить эффект от воздействия, если он существует. Способность обнаружения обычно называют *мощностью* или *чувствительностью*. Один из способов улучшить чувствительность – уменьшить дисперсию.

Список литературы

1. *Маккинли У.* Python и анализ данных, 2015. – 482 с.
2. *Кохави Р.* Доверительное A/B-тестирование. Практическое руководство по контролируемым экспериментам, 2021. – 298 с.