**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**Федеральное государственное автономное образовательное**

**учреждение высшего образования**

**«Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»**

**Институт экономики и предпринимательства**

**Кафедра математического моделирования экономических процессов**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА**

**НАПРАВЛЕНИЕ 38.03.05 «БИЗНЕС-ИНФОРМАТИКА»**

**А/В тестирование: статистический анализ и прогнозирование результатов**

**Зав. кафедрой ММЭП ИЭП ННГУ**

**д.ф.-м.н., профессор**

**Кузнецов Ю.А.**

**Исполнитель:**

**студентка 4 курса**

**Коковина А. В.**

**Научный руководитель:**

**доцент кафедры ММЭП ИЭП ННГУ,**

**к.ф.-м.н.**

**Капитанова О.В.**

**Нижний Новгород, 2023**

# Аннотация

Данная работа посвящена А/В тестированию, как способу сравнения результатов различных маркетинговых стратегий. В настоящее время подобные вопросы представляются весьма актуальными, в связи с тем, что различные методики могут для разных бизнес-направлений работать неодинаково. Поэтому, анализ конкретных стратегий для своего бизнеса позволяет повысить ключевые показатели эффективности, а А/В тестирование сделает результаты научно-обоснованными и достоверными.

Целью выпускной квалификационной работы являлось изучение существующих подходов к А/В тестированию и реализация их для статистических данных с помощью Python.

Содержание работы предусматривает введение, 3 главы, заключение, список использованной литературы и 5 приложений. В первой главе работы рассмотрены теоретические вопросы, связанные с проведением самого А/В тестирования, вторая глава посвящена вопросам подготовки данных. В третьей главе представлено практическое исследование использования различных методов А/В тестирования на синтетических и реальных данных с помощью Python.

Объем работы – 55 страниц. Работа содержит 57 рисунков. Список литературы включает 6 источников.

**Содержание**

[Введение 4](#_Toc135582755)

[Глава 1. A/B тестирование: методы, подходы и особенности проведения 5](#_Toc135582756)

[1.1. A/B тестирование. Общее понятие 5](#_Toc135582757)

[1.2. Методы проведения A/B тестирования 7](#_Toc135582758)

[1.3. Частотный метод проведения A/B тестирования 8](#_Toc135582759)

[1.4. Байесовский метод в A/B тестировании 14](#_Toc135582760)

[Глава 2. Методы предобработки данных 20](#_Toc135582761)

[2.1. Методы обнаружения выбросов в анализируемых данных 20](#_Toc135582762)

[2.2. Тесты на определение нормальности распределения 20](#_Toc135582763)

[2.3. Особенности проведения тестов для не нормально распределённых данных 23](#_Toc135582764)

[Глава 3. A/B тестирование, практическая реализация программными средствами 25](#_Toc135582765)

[3.1. Байесовский метод 25](#_Toc135582766)

[3.2. Частотный метод 34](#_Toc135582767)

[3.3. А/В тестирование: реальный данные 43](#_Toc135582768)

[Заключение 54](#_Toc135582769)

[Список литературы 55](#_Toc135582770)

# Введение

В современном мире бизнес-аналитика и маркетинговые исследования становятся все более значимыми для успешной работы компаний в условиях жесткой конкуренции. Одним из ключевых инструментов, используемых в данном контексте, является A/B тестирование, которое позволяет оценить эффективность различных маркетинговых и бизнес-стратегий.

А/B тестирование является методом экспериментального дизайна, который позволяет сравнить две или более версии одного и того же элемента веб-страницы, приложения, продукта или сервиса. Тестирование позволяет определить, какие изменения в дизайне или функциональности приводят к улучшению метрик конверсии и продаж.

Однако, для достижения достоверных результатов необходимо использовать правильные методы статистического анализа. Важным этапом A/B тестирования является оценка статистической значимости полученных результатов и прогнозирование их будущих значений. Неправильное применение статистических методов может привести к неверной интерпретации результатов тестирования и неправильному принятию решений.

В связи с этим, выбранная тема представляется весьма актуальной и важной для бизнес-аналитиков и маркетологов. В данном исследовании будет рассмотрено использование различных методов статистического анализа для A/B тестирования, а также методы прогнозирования будущих результатов. Результаты исследования могут быть полезными для принятия правильных бизнес-решений и улучшения эффективности маркетинговых кампаний и продуктов.

В первой главе работы рассмотрены теоретические вопросы, связанные с проведением самого А/В тестирования, вторая глава посвящена вопросам подготовки данных. В третьей главе представлено практическое исследование использования различных методов А/В тестирования на синтетических и реальных данных с помощью Python.

# Глава 1. A/B тестирование: методы, подходы и особенности проведения

# 1.1. A/B тестирование. Общее понятие

А\В тестирование (или split testing) – это методика экспериментального исследования, которая используется для определения наиболее эффективной стратегии, которая приводит к наибольшему количеству желаемых результатов.

В А\В тестировании исследователи случайным образом разделяют группу испытуемых на две части: контрольную группу (А) и экспериментальную группу (B). Затем в каждой группе применяются различные изменения в интересующих исследователей переменных, например, дизайн сайта, текст рекламного объявления, цвет кнопки и т.д. Исследователи затем сравнивают результаты между группами, чтобы определить, какие изменения были наиболее эффективны.

Для проведения А\В тестирования применяются различные статистические инструменты. Один из самых распространенных инструментов – t-тест Стьюдента, который используется для проверки статистической значимости различий между двумя группами данных. Другой распространенный инструмент – доверительный интервал, который позволяет исследователям определить, насколько точны результаты их тестирования.

В дополнение к этому А\В тестирование может включать анализ ковариации, который используется для контроля за нежелательными факторами, которые могут повлиять на результаты тестирования. Также могут применяться другие методы, такие как регрессионный анализ и анализ временных рядов, в зависимости от конкретных условий исследования.

Можно выделить значимые преимущества А/B тестирования:

* Увеличение конверсии: А/B тестирование помогает улучшить эффективность веб-сайта или приложения, что может увеличить конверсию и выручку.
* Оптимизация пользовательского опыта: тестирование позволяет исследовать, как пользователи взаимодействуют с продуктом и улучшить его интерфейс и удобство использования.
* Надежность: А/B тестирование является наиболее надежным методом тестирования и может предоставить точную статистическую информацию о том, какой вариант является наиболее эффективным.
* Лучшее принятие решений: А/B тестирование помогает бизнесу принимать более осознанные и обоснованные решения, основанные на результатах тестирования.

Несмотря на это существуют некоторые недостатки А/B тестирования:

* Ограниченность данных: А/B тестирование может ограничивать данные, доступные для анализа, что может ограничить количество переменных, которые можно учитывать в тестировании.
* Затратность: А/B тестирование может быть дорогим, особенно если планируется проводить множество тестов.
* Сложность внедрения: Реализация тестирования может потребовать значительных технических усилий.

Существуют различные подходы к А/B тестированию:

* A/A тестирование: тестирование, при котором используется одинаковый контрольный вариант, чтобы проверить, насколько точны результаты.
* A/B/n тестирование: тестирование, при котором тестируется более двух вариантов. Этот подход позволяет проводить более широкое тестирование и сравнивать различные элементы дизайна.
* Мультитестирование: подход позволяет проводить несколько тестов одновременно, чтобы сократить время и стоимость.

На практике выделяют 7 этапов проведения такого тестирования:

1. Определение целей и метрик.

Первым шагом в проведении A/B тестирования является определение целей и метрик. Цели должны быть ясными и конкретными, а метрики должны быть выбраны таким образом, чтобы отражать эти цели. Например, для увеличения конверсии, метрикой может быть количество людей, которые совершили покупку на сайте.

1. Создание гипотезы.

Создание гипотезы – это следующий шаг в проведении A/B тестирования. Гипотеза должна основываться на определённых ранее целях и метриках. Она должна быть конкретной и проверяемой. Например, гипотеза может звучать так: «Изменение цвета кнопки «купить» на красный увеличит конверсию на 10%».

1. Реализация.

После определения гипотезы следует создание вариантов. Варианты должны отличаться только одним элементом. Например, если вы тестируете цвет кнопки, то остальные элементы должны оставаться неизменными.

1. Разделение трафика.

Для проведения A/B тестирования необходимо разделить трафик на две группы: контрольную и тестовую. Контрольная группа будет видеть старый вариант, а тестовая - новый вариант.

1. Проведение теста.

Проведение теста – это самый важный шаг в A/B тестировании. Тест должен продолжаться в течение определенного периода времени, чтобы получить надежные результаты. Обычно этот период составляет не менее недели.

1. Анализ результатов.

После окончания теста нужно проанализировать результаты. Метрики, которые были выбраны, должны быть измерены и сравнены между контрольной и тестовой группами. Результаты должны быть статистически значимыми.

1. Принятие решения.

На основе результатов теста нужно принимать гипотезу о выигравшем варианте или гипотезу о том, что различий нет.

# 1.2. Методы проведения A/B тестирования

Широкое применение получили 2 метода проведения А\В тестирования: частотный и байесовский.

Частотный метод использует статистические методы для проверки гипотезы о различии между двумя группами. Он основан на определении уровня значимости (обычно 0,05) и нахождении P-значения (вероятность получения разницы между двумя группами, равной или более экстремальной, чем фактическая разница), при условии, что нулевая гипотеза (что разницы между группами нет) верна. Если P-значение меньше уровня значимости, нулевая гипотеза отвергается в пользу альтернативной гипотезы (что разница между группами существенна).

Байесовский метод, в свою очередь, использует байесовскую статистику для оценки вероятности того, что различия между группами статистически значимы. Он основан на априорных предположениях о вероятности различий между группами, а затем обновляет эти предположения на основе полученных данных, чтобы получить апостериорные распределения вероятностей.

Сравнение двух методов:

* Интерпретация результатов: Частотный метод даёт ответ на вопрос, существует ли статистически значимая разница между группами или нет. Байесовский подход же дает ответ на вопрос, насколько велика вероятность того, что существует статистически значимая разница между группами.
* Обработка априорной информации: Частотный подход не использует априорную информацию, а байесовский подход основан на ее использовании. В байесовском подходе априорные предположения могут быть сильными, если есть доступ к дополнительной информации.
* Устойчивость к выбросам: Байесовский подход более устойчив к выбросам, т.к. он учитывает априорную информацию и не полагается только на статистические методы.

# 1.3. Частотный метод проведения A/B тестирования

Частотный метод проведения А/В тестирования заключается в том, что на выборке пользователей случайным образом назначаются разные варианты тестовых групп и контрольных групп. Вариант А является контрольным, а вариант В является тестовым. После назначения групп идет наблюдение за поведением пользователей, чтобы оценить, как каждый вариант влияет на конечный результат.

Частотный метод А/В тестирования является статистическим методом, основанным на анализе вероятностных распределений. Он позволяет контролировать влияние случайных факторов на результаты тестирования и увеличивает достоверность выводов.

Для проведения А\В тестирования с помощью частотного метода необходимо собрать данные о поведении пользователей в обеих группах, например, количество кликов на определенную кнопку на веб-сайте или количество продаж после изменения цвета кнопки. Затем, используя статистические методы, сравниваются результаты между группами, чтобы определить, есть ли статистически значимые различия между ними.

Однако, существуют ограничения для данных, которые могут повлиять на точность результатов А\В тестирования. Во-первых, недостаточное количество данных может привести к неправильному анализу и неверным выводам. Во-вторых, нерепрезентативная выборка может привести к искажению результатов тестирования, например, если в экспериментальной группе больше молодых пользователей, а в контрольной группе - более старших, то это может исказить результаты тестирования. Наконец, некорректное разделение на группы, например, если пользователи случайно попадают в определенную группу, может также привести к ошибочным выводам из-за возможных искажений.

Некоторые методы, наиболее часто применяемые в рамках данного подхода будут описаны ниже, а так же в Главе 2.

Если обе части выборки (тестовая и контрольная) имеют нормальное распределение принято применять параметрический тест Левена на равенство дисперсий.

В статистике Тест Левена - это статистический вывод, используемый для оценки равенства отклонения для переменной, рассчитанной для двух или более групп. Некоторые общие статистические процедуры предполагают, что дисперсия совокупностей, из которых взяты разные выборки, равны. Тест Левена оценивает это предположение. Он проверяет нулевую гипотезу о том, что дисперсии генеральной совокупности равны (это называется однородностью дисперсии или гомоскедастичностью ). Если результирующее p-значение критерия Левена меньше некоторого уровня значимости (обычно 0,05), полученные различия в дисперсиях выборки маловероятны на основе случайной выборки из генеральной совокупности с равными дисперсиями. Таким образом, нулевая гипотеза о равных дисперсиях отклоняется и делается вывод о различии дисперсий в генеральной совокупности.

В случае если дисперсии равны, предполагается применить стандартный независимый t-критерий с двумя выборками, иначе предполагается провести Т-критерий Уэлча - этот критерий предполагает, что обе группы данных взяты из совокупностей, которые следуют нормальному распределению, но не предполагает, что эти две совокупности имеют одинаковую дисперсию.

В рамках частотного подхода для оценки статистической разницы между выборками весьма часто применяется метод «бутстрэп». Бутстрэп – это оценка структурных характеристик генеральной совокупности на основе перевыборки (resampling) из выборки. Иными словами, перевыборка по отношению к выборке рассматривается как выборка по отношению к генеральной совокупности. В A/B тестировании бутстрэп можно использовать для оценки статистических характеристик различий между группами, таких как среднее значение или медиана.

Идея бутстрэп заключается в том, что из исходной выборки многократно выбираются подвыборки с повторениями (бутстрэп-выборки), которые имеют тот же размер, что и исходная выборка. На каждой бутстрэп-выборке вычисляются статистические характеристики, например, среднее значение, и собираются в распределение. Это распределение характеризует неопределенность в оценке статистических характеристик на основе конкретной выборки.

Бутстрэп можно использовать для определения значимости различий между группами в A/B тестировании. Например, при оценке различия между средними значениями конверсии в двух группах, можно применить бутстрэп для вычисления доверительных интервалов для различия между средними значениями.

Одним из преимуществ использования бутстрэп в A/B тестировании является его способность учитывать неопределенность, которая может возникнуть при работе с малыми выборками. Кроме того, бутстрэп является непараметрическим методом, который не требует знания распределения выборки, что может быть полезно при работе с выборками, которые не следуют нормальному распределению.

Однако, при использовании бутстрэп необходимо учитывать, что этот метод также имеет некоторые ограничения. Например, он может быть чувствителен к выбросам в выборке, поэтому важно проводить анализ выбросов перед его использованием. Кроме того, бутстрэп может потребовать вычислительных ресурсов, особенно если выборка очень большая.

В качестве оценки функции распределения можно использовать эмпирическую функцию распределения (ЭФР). ЭФР является несмещенной оценкой и сходится к истинной ФР при увеличении размера выборки.

Функцией распределения случайной величины Х называется функция задаваемая формулой:

Дана выборка независимых одинаково распределенных случайных величин . Эмпирическая функция распределения имеет вид:

На рисунке 1.1 видно, что при увеличении размера выборки ЭФР лучше приближает истинную функцию распределения. Если увеличить размер выборки до нескольких тысяч, то ЭФР визуально будет сложно отличить от истинной функции распределения.

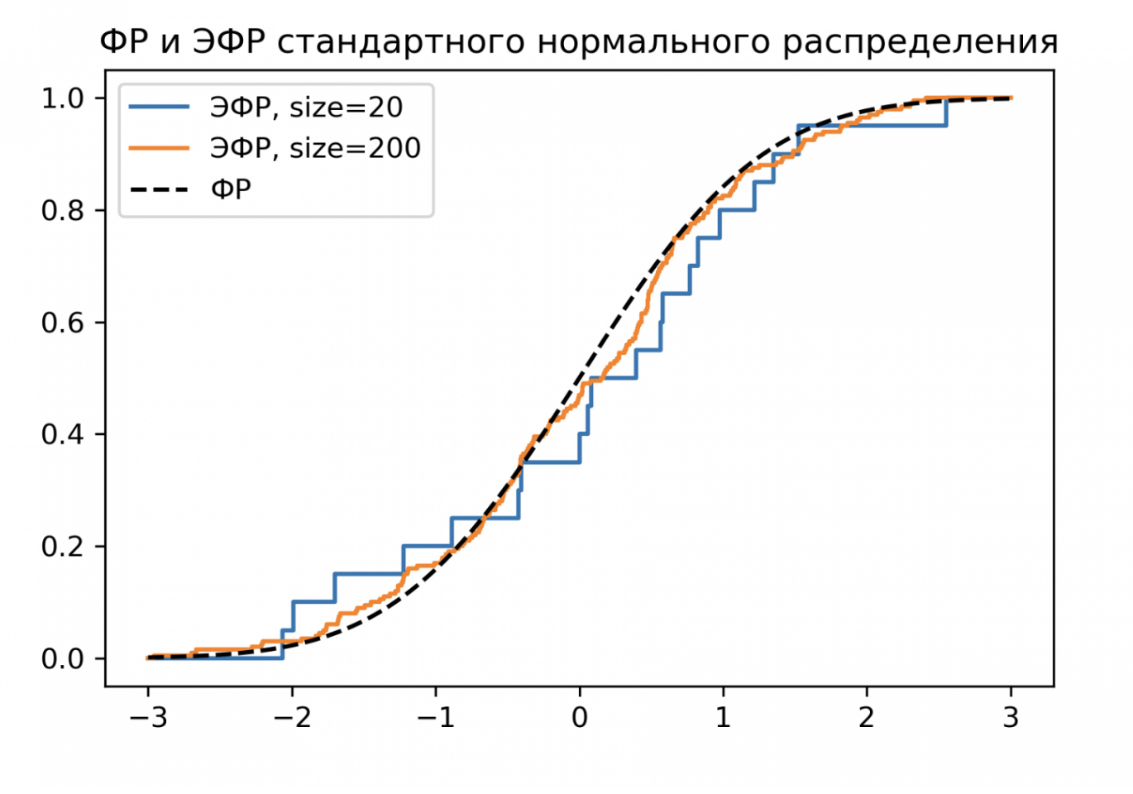


Рис. 1.1. ЭФР для данных разного размера из стандартного нормального распределения

С помощью этого метода можно оценивать стандартное отклонение оценки статистики. Зная стандартное отклонение, можно интуитивно понять, насколько достоверны полученные результаты. Для получения более точных результатов можно построить доверительный интервал.

Доверительный интервал (ДИ) – это интервал, который покрывает оцениваемый параметр с заданной вероятностью.

Доверительным интервалом с доверительной вероятностью 1-*a* для параметра называется интервал где a и b такие функции выборки, что

В научном контексте вероятность доверительного интервала следует понимать как вероятность многократного проведения эксперимента, дающее в 95% случаев истинное значение параметра, располагающееся в доверительном интервале. Однако следует отметить, что в строгом смысле доверительный интервал является случайным вектором. В реальности предстоит иметь дело с конкретной реализацией доверительного интервала, и говорить о вероятности отличной от 0 или 1 нецелесообразно, так как истинное значение параметра либо принадлежит интервалу, либо не принадлежит. Однако для удобства используем термин «доверительный интервал» вместо «реализация доверительного интервала», также как часто используем термин «выборка» вместо «реализация выборки».

Рассмотрим 95% доверительный интервал. Если данных много, то, независимо от распределения исходных данных (предполагается, что дисперсия существует и не равна нулю), по центральной предельной теореме среднее будет распределено нормально. Для нормально распределённых статистик ДИ (рис. 1.2) можно вычислить по формуле:

-квантиль стандартного нормального распределения, *а* – уровень значимости.

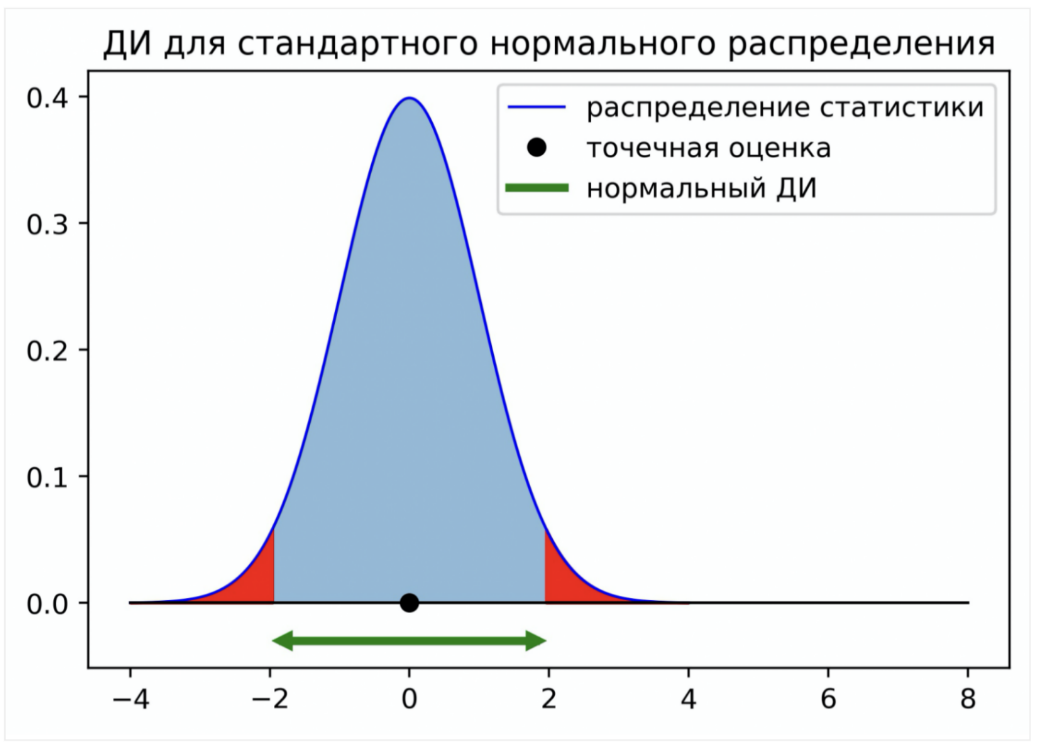


Рис. 1.2. Демонстрация доверительного интервала для стандартного нормального распределения

Для построения ДИ квантиля будем использовать бутстрэп. Это можно сделать тремя способами.

Первый способ:

* оценить значение квантиля по исходным данным;
* с помощью Бутстрэп оценить стандартное отклонение оценки квантиля;
* по формуле вычислить ДИ

*pe* – точечная оценка

Такой ДИ называется нормальным доверительным интервалом.

Использование нормального доверительного интервала эффективно, если статистика распределена близко к нормальному распределению. Однако, в случае несимметричного распределения, применение нормального доверительного интервала может привести к искажению результатов. На приведенном ниже рисунке 1.3 иллюстрируется ситуация, когда граница доверительного интервала находится за пределами минимального значения распределения.



Рис. 1.3. Демонстрация доверительного интервала для логнормального распределения

В случае несимметричных распределений можно использовать процентильный доверительный интервал (рис. 1.4). Чтобы построить процентильный ДИ, нужно отрезать с каждой стороны по a/2 площади распределения. Для 95% ДИ нужно отрезать по 2.5%. На практике для вычисления границ ДИ нужно оценить квантиль *Qa/2* и по *Q(a-1)/2* значениям статистик, полученных с помощью бутстрэпа. Доверительный интервал будет иметь следующий вид

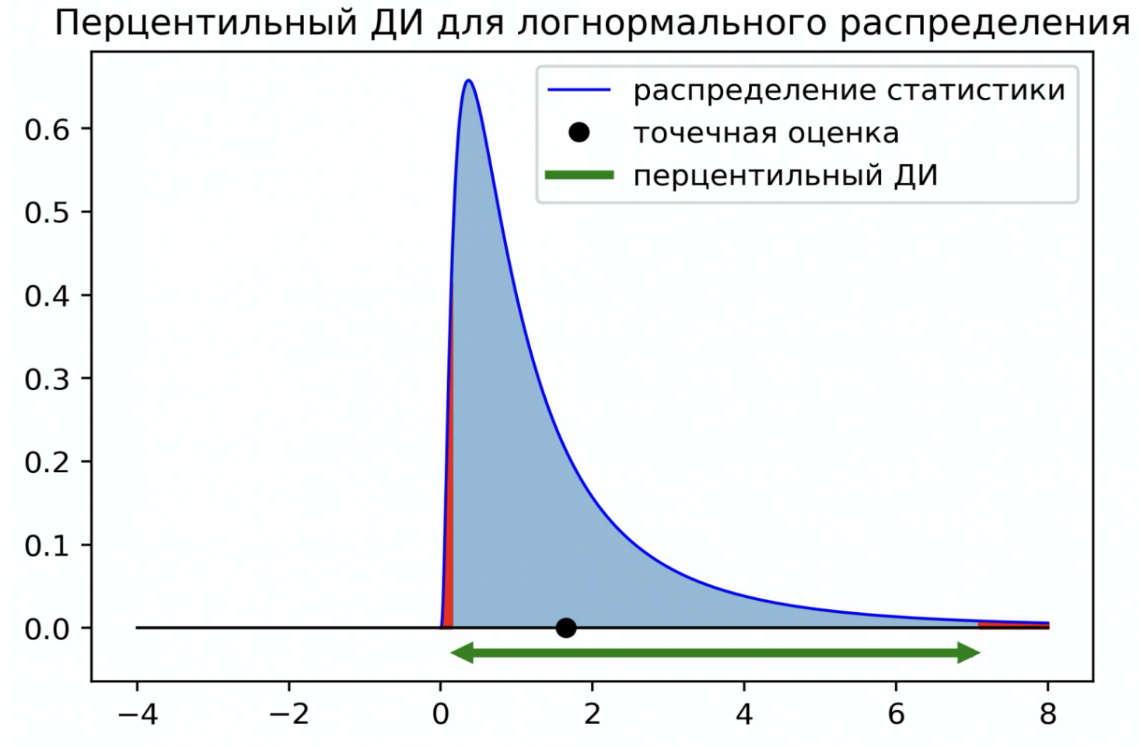


Рис. 1.4. Демонстрация доверительного интервала для логнормального распределения

Существует ещё один вариант - центральный доверительный интервал (рис. 1.5). Его границы равны:

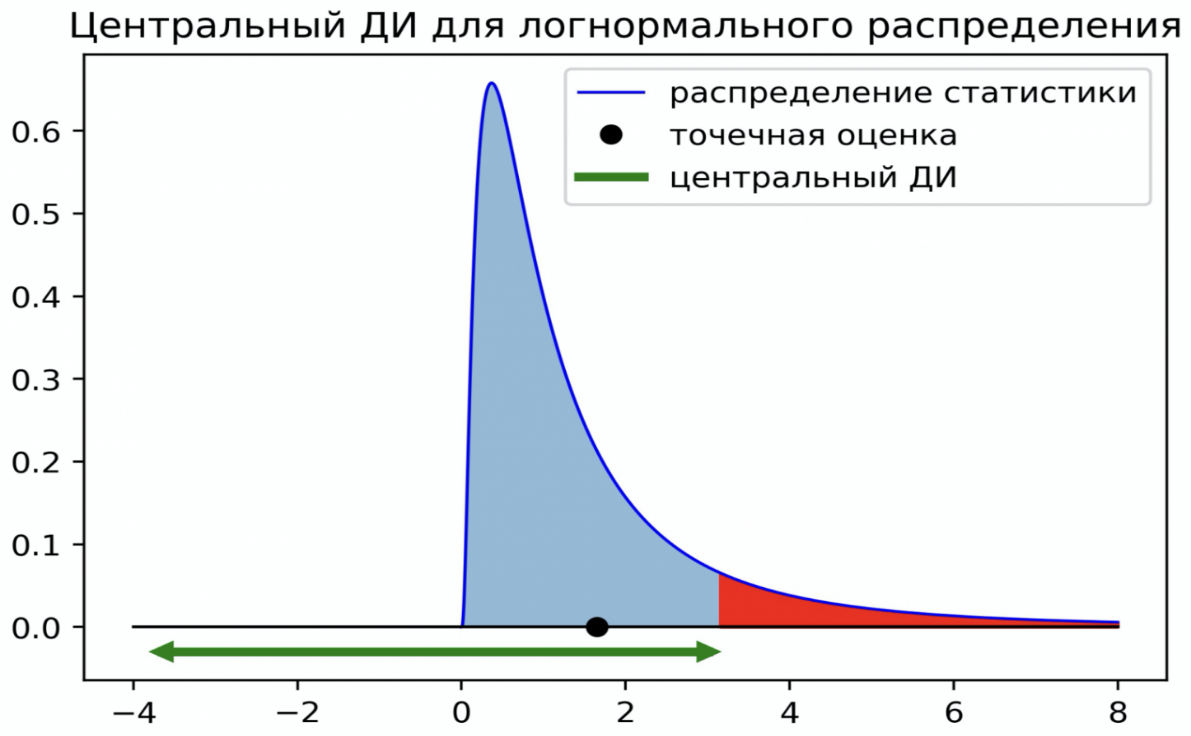


Рис. 1.5. Демонстрация доверительного интервала для логнормального распределения

Центральный доверительный интервал для логнормального распределения смещён влево относительно нуля, что связано с тем, что логнормальное распределение имеет ненулевую плотность вероятности только для положительных значений. Этот сдвиг можно рассматривать как проявление осторожности в случае, когда не неизвестна полная информация о распределении и возможны значения меньше нуля.

С помощью доверительного интервала можно не только получать дополнительную информацию о значении метрики, но и проверять статистические гипотезы. Чтобы проверить гипотезу о равенстве квантилей на уровне значимости 5% достаточно построить 95% доверительный интервал для разности квантилей между группами. Если ноль находится вне доверительного интервала, то отличия статистически значимы, иначе нет.

# 1.4. Байесовский метод в A/B тестировании

Байесовский метод в A/B тестировании – это статистический подход, который позволяет учитывать вероятность правильности гипотезы в процессе тестирования. Этот метод основан на теореме Байеса, которая позволяет обновлять вероятность события на основе новых данных.

где:

P(A) - априорная вероятность

P(A|B) - вероятность гипотезы А при наступлении события В

P(B|A) - вероятность наступления события В при истинности гипотезы А

P(B) - полная вероятность наступления события В

При использовании байесовского метода в A/B тестировании, каждый раз, когда появляются новые данные, необходимо пересчитывать вероятность правильности гипотезы. Вместо того, чтобы просто выбрать победителя, считается вероятность того, что один вариант лучше другого.

Например, при проведении A/B тестирования на сайте, чтобы определить, какой заголовок страницы работает лучше. Начинать необходимо с двух гипотез: «Заголовок А лучше» и «Заголовок В лучше». Далее разделить трафик на две группы и тестировать оба заголовка.

Каждый раз, при получении новых данных о конверсии, происходит обновление вероятности на основе теоремы Байеса. Если данные говорят, что вероятность того, что заголовок А лучше, можно сделать вывод, о том, что заголовок А действительно лучше.

Преимущество байесовского подхода заключается в том, что появляется возможность получить более точные результаты с меньшим количеством данных, чем в случае с традиционными статистическими методами. Кроме того, байесовский метод позволяет учитывать априорные знания, что может быть полезно, когда рассматриваются новые тесты или маленькие выборки.

При использовании байесовского метода в A/B тестировании, есть определенные требования к данным, которые нужно учитывать:

* Данные должны быть достаточно чистыми и точными. Необходимо убедиться, что данные, точно соответствуют цели измерения, и что они не содержат ошибок или искажений. В противном случае, велика возможность получить неверные результаты и неправильно принять решение.
* Данные должны быть достаточно большими. Чтобы получить достоверные результаты, нужно собрать достаточное количество данных, чтобы иметь достаточную степень уверенности в нашем выводе. В противном случае, есть риск получить слишком широкие доверительные интервалы и неправильно оценить результаты тестирования.
* Данные должны быть случайными и независимыми. Важно, чтобы данные, используемые для тестирования, были случайными и независимыми друг от друга. Это поможет избежать искажений в результате и гарантировать, что результаты тестирования будут честными и точными.
* Данные должны быть представлены в правильном формате. Для использования байесовского метода в A/B тестировании, данные должны быть представлены в правильном формате, который позволяет извлекать статистические выводы из них. Например, данные должны быть представлены в виде числовых значений (как конверсионные данные) или долей (как отношение количества кликов к общему числу посещений страницы).

В целом, байесовский подход в A/B тестировании может дать более точные результаты, чем традиционные статистические методы. Но при использовании этого метода, необходимо учитывать специфические требования к данным и необходимо грамотно подходить к их сбору и анализу.

Основные шаги для проведения байесовского A/B-тестирования включают в себя три этапа:

1. Сбор данных для эксперимента;
2. Сравнение различных вариантов, применяя теорему Байеса;
3. Принятие решения о том, достигнут ли эксперимент статистически значимых результатов и может ли быть остановлен.

Эти три шага проиллюстрированы в блок-схеме (рис. 1.6), которая позволяет получить более полное представление о различных решениях, которые должен принять специалист.

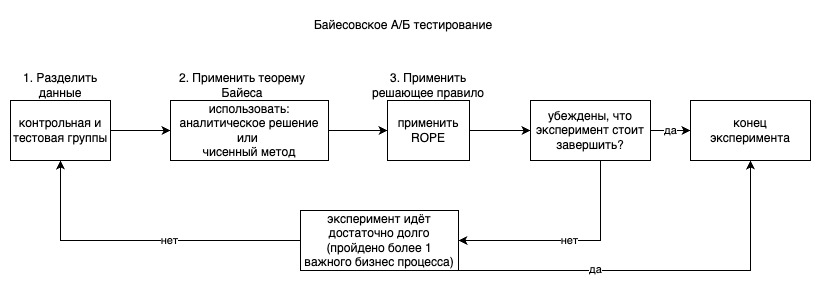


Рис. 1.6 Схема проведения байесовского А/В теста.

Перепишем теорему Байеса для дальнейшего использования:

P(H|d) – апостериорная вероятность. Это вероятность того, что какая-то конкретная гипотеза верна после сбора и добавления новых данных. Рассуждая в терминах А\В тестирования, она выражает вероятность получить весь набор возможных исходов теста. Например, коэффициент конверсии для A лучше, чем для B, на +0,1%, +0,2%, -0,1%, -0,2%, … и каждый промежуточный пункт и далее. Множество всех возможных исходов бесконечно, а это значит, что вероятностное поле исходов будет представлено непрерывным распределением.

P(d|H) – функция правдоподобия. Она показывает, насколько вероятно наблюдать собранные данные для заданной гипотезе модели, представлена непрерывным распределением.

P(H) – априорное распределение. Оно кодирует информацию о модели до появления реальных данных. Например, предварительная информация о том, что средний коэффициент конверсии для A должен быть сосредоточен вокруг 10% со стандартным отклонением в 1%. Тогда предварительная оценка для A будет представлена соответствующим нормальным распределением. Поскольку нет информации о B, лучше всего предположить, что B ведет себя так же, как A.

P(d) – это мера того, насколько хорошо модель соответствует данным (насколько хороша модель). Для непрерывных распределений, можно оценить:

При нахождении P(H|d) обычно легко вычислить числитель теоремы Байеса, но трудно оценить P(d), только зная его нормирующую константу, можно сделать апостериорное распределение фактическим распределением вероятности (интегрируемым в единицу), которое затем можно использовать для вычисления любых других величин (обычно называемых «моментами» функции распределения). Обычно можно вычислить апостериорное распределение - включая его нормирующую константу - только с помощью численных расчетов. В частности, обычно используются методы Марковской цепи Монте-Карло. Однако для некоторых моделей определенного типа оказывается, что апостериорное распределение может быть вычислено аналитически.

Аналитический метод: для того, чтобы решить эту проблему аналитически, необходимо тщательно выбрать предварительные распределения в зависимости от формы функции правдоподобия. Для стандартного A/B-теста «коэффициент конверсии» функция правдоподобия имеет очень простую форму: распределение Бернулли (экспериментальные данные должны содержать два массива, один для контрольной группы и один для тестовой. Каждая запись массива будет представлять, перешел ли определенный пользователь (значение=1) или нет (значение=0)).

Теперь назовем два массива двоичных значений как dA для группы А (контрольной) dB для группы Б (тестовой). Кроме того, допустим, что Контрольная группа имеет вероятность pA обращения, в то время как в тестовой группе вероятность pB. Тогда совместная вероятность P(d|H) которая входит в теорему Байеса, может быть оценена как:

Для конкретных распределений правдоподобия можно найти конкретные формы априорноого распределения, которые делают последующую оценку теоремы Байеса очень простой. Для целей A/B тестирования, учитывая выражение совместной вероятности P(d|H), описанной выше эта «специфическая форма» представлена бета-распределением, которое является непрерывным распределением, определенным как функция с параметрами >0, >0

Г – гамма-функция.

При выборе предварительного уравнения (1.5) получается, что апостериорное распределение также является бета-распределением. Поскольку предшествующее и последующее распределения принадлежат к одному и тому же семейству бета-распределений, предшествующее распределение из уравнения (1.5) называется сопряженным априорным распределением функции правдоподобия.

Для того чтобы определить «победившую» версию необходимо понять когда нужно остановить эксперимент и как оценить вероятность, для этих целей будет рассмотрен метод практического эквивалента (ROPE)

Ведём некоторые понятия:

Интервал высокой задней плотности (HPD) – область HPD для параметра θ является подмножеством C всех возможных значений параметра θ определяемое:

где наибольшее k такое, что:

Предположим, что мы ищем 95%-ную HPD, что означает, что мы принимаем α=0,95. Чтобы определить 95% HPD для распределения на рисунке 1.7 ниже, следует поступить следующим образом: начинаем с горизонтальной линии над максимумом, затем начинаем опускать ее и измеряем интеграл (площадь) ℙ(θ|d) для всех значений, которые находятся выше этой горизонтальной линии. В конечном итоге приходим к значению, для которого интеграл равен 0,95. Это и есть значение k, которое определяет HPD.

ROPE – это то, к чему исследователь должен прийти до проведения эксперимента в качестве ответа на следующий вопрос: «При каких обстоятельствах можно сказать, что между двумя группами A и B фактически нет разницы?» Это произойдет, когда большая часть «разницы» между группами A и B лежит в пределах ROPE. ROPE определяется на основании размера эффекта. Размер эффекта ES – это мера разницы в среднем между двумя группами, нормированная на стандартное отклонение.

где μA и μB выборочные средние для групп A и B, соответственно, и σ неопределенное значение для стандартного отклонения. Обычно стандартное отклонение разности средних оценивается как:

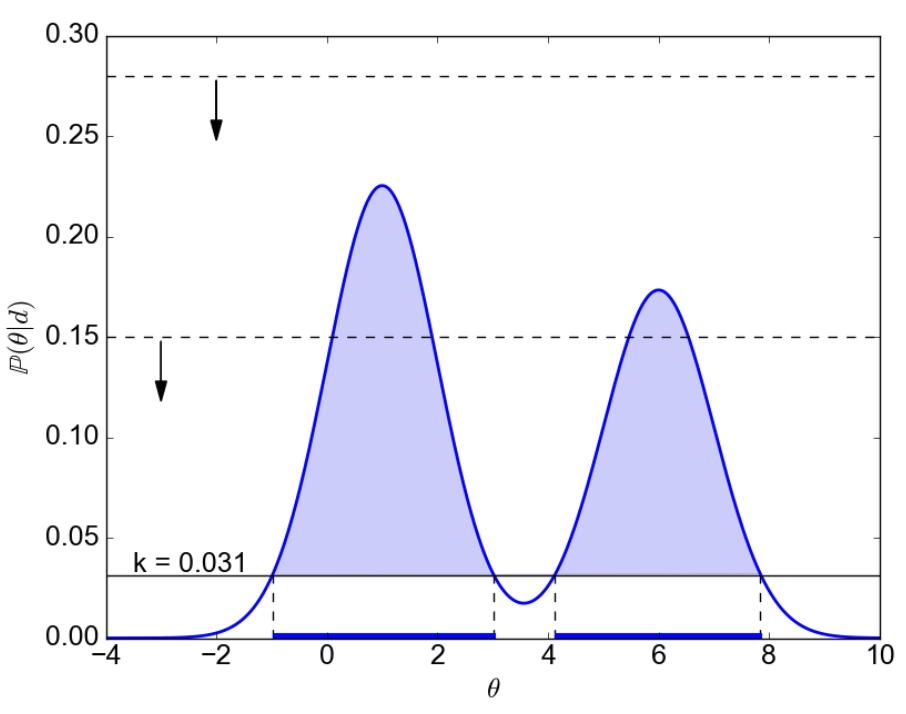


Рис. 1.7. Поиск HPD

Если ES = ±0,1, это означает, что разница в средних значениях между группами А и В составляет всего 10% от значения стандартного отклонения. Довольно маленькое значение и что-то, что вполне может соответствовать определению «незначительно отличающихся» между двумя группами.

# Глава 2. Методы предобработки данных

# 2.1. Методы обнаружения выбросов в анализируемых данных

Определение выбросов в выборке – это один из этапов подготовки данных, который позволяет исключить ошибочные или неправильные значения, которые могут повлиять на результаты анализа данных. Ниже перечислены некоторые методы определения выбросов:

1. Метод межквартильного расстояния (IQR): данный метод основывается на разнице между верхним (75-й квартиль) и нижним (25-й квартиль) квартилями. Любое значение, которое находится за пределами IQR умноженного на 1,5 (или 3) является выбросом.
2. Z-оценка: данный метод использует стандартное отклонение для определения выбросов. Любое значение, которое находится за пределами Z-оценки более чем на 3 (или 2) стандартных отклонения является выбросом.

где X — значение точки данных; μ — среднее значение; σ — стандартное отклонение.

1. Box plot: это графический метод, который показывает распределение данных в выборке и позволяет определить выбросы. Любое значение, которое находится за пределами "усов" (между верхним и нижним квартилями), является выбросом.

# 2.2. Тесты на определение нормальности распределения

Для того чтобы определить, есть ли статистически значимые различия между группами в частотном методе, в А/В тестировании используются различные инструменты. Для того чтобы выбрать их для данного конкретного случая, необходимо определить является ли выборка нормально распределённой. Для этого используют статистические тесты на нормальность.

1. Тест Шапиро-Уилка

Критерий Шапиро-Уилка основан на оптимальной линейной несмещённой оценке дисперсии к её обычной оценке методом максимального правдоподобия. Этот тест является одним из наиболее распространенных тестов на нормальность и может использоваться для выборок любого размера. Однако он может быть неэффективен для очень больших выборок (более 5000 наблюдений). Также, если выборка содержит выбросы, то тест может быть нечувствительным и дать неверные результаты.

Статистика критерия имеет вид:

где:

1. Критерий согласия Андерсона — Дарлинга

Классический непараметрический критерий согласия Андерсона — Дарлинга предназначен для проверки простых гипотез о принадлежности анализируемой выборки полностью известному закону (известным вектором параметров теоретического закона. Этот тест может быть использован для выборок любого размера, и он более чувствителен к выбросам, чем тест Шапиро-Уилка. Однако, он также может быть неэффективен для очень больших выборок.

Значение статистики Андерсона-Дарлинга S\* вычисляется по формуле:

Значение вероятности вычисляется по функции распределения

Критические значения критерия Sa при заданном a могут быть взяты из таблицы критических значений

Гипотеза H0 не отвергается, если для вычисленного по выборке значения статистики S\*

1. Тест Лиллиефорса

Третьим тестом на нормальность, который будет рассмотрен в данной работе, является тест Лиллиефорса. Этот тест может быть использован для выборок любого размера, и он более чувствителен к выбросам, чем тест Колмогорова-Смирнова. Однако, он менее чувствителен, чем тесты Шапиро-Уилка и Андерсона-Дарлинга. Используется для проверки нулевой гипотезы о том, что выборка распределена по нормальному закону для случая, когда параметры нормального распределения (математическое ожидание и дисперсия) априори неизвестны.

Проверка гипотезы проводится следующим образом:

* Оценивается выборочное среднее и дисперсия;
* Так же как при использовании критерия Колмогорова, находится максимальное отклонение между выборочной и теоретической интегральными функциями распределения;
* Принимается решение, является ли статистически значимым наблюдаемое отклонение выборочной функции распределения от теоретической. В случае положительного ответа, нулевая гипотеза отвергается.

Важным аспектом при выборе теста на нормальность является мощность теста.

Для анализа нормальности выборки размером больше 50000 рекомендуется использовать тесты на нормальность, которые основаны на критериях эксцесса и асимметрии, так как они обладают более высокой мощностью для больших выборок.

Один из таких тестов – это тест Харке-Бера, который использует оценки эксцесса и асимметрии выборки для проверки её нормальности. это статистический тест, проверяющий ошибки наблюдений на нормальность посредством сверки их третьего момента (асимметрия) и четвёртого момента (эксцесс) с моментами нормального распределения.

Другой тест на основе критериев эксцесса и асимметрии – это тест Д'Агостино-Пирсона. Он основан на оценках коэффициентов эксцесса и асимметрии и использует эмпирические распределения для вычисления p-значения.

Важно помнить, что результаты тестов на нормальность не должны рассматриваться изолированно, а всегда должны быть рассмотрены в контексте конкретной задачи и других статистических методов, используемых для анализа данных.

# 2.3. Особенности проведения тестов для не нормально распределённых данных

Обычно при А\В тестировании используются статистические методы, которые предполагают, что данные распределены нормально. Однако, в некоторых случаях распределение данных может отличаться от нормального, что может привести к неправильной интерпретации результатов тестирования.

К примеру, когда данные имеют скошенное распределение, как например в случае с «тяжелыми хвостами», то применение методов, основанных на нормальном распределении, может быть неверным. Также, если данные имеют выбросы, это может привести к искажению результатов.

Одним из подходов к работе с данными, распределение которых не является нормальным, является использование бутстрэпа - метод, который позволяет создать множество выборок данных и сделать выводы на основе этих выборок.

Другой подход - использование непараметрических тестов, таких как тест Уилкоксона-Манна-Уитни. Для того чтобы его провести, необходимо:

1. Составить единый ранжированный ряд из обеих сопоставляемых выборок, расставив их элементы по степени нарастания признака и приписав меньшему значению меньший ранг (при наличии повторяющихся элементов в выборке использовать средний ранг). Общее количество рангов получится равным где и количество элементов в первой и второй выборке соответственно

2. Разделить единый ранжированный ряд на два, состоящих соответственно из единиц первой и второй выборок. Подсчитать отдельно сумму рангов, пришедшихся на долю элементов первой выборки , и отдельно — на долю элементов второй выборки, затем вычислить:

Если вычислено верно, тогда:

3.Определить значение U-статистики Манна-Уитни по формуле

4. По таблице для избранного [уровня статистической значимости](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%B7%D0%BD%D0%B0%D1%87%D0%B8%D0%BC%D0%BE%D1%81%D1%82%D1%8C) определить критическое значение критерия для данных n1 и n2. Если полученное значение U меньше табличного или равно ему, то признается наличие существенного различия между уровнем признака в рассматриваемых выборках (принимается [альтернативная гипотеза](https://ru.wikipedia.org/w/index.php?title=%D0%90%D0%BB%D1%8C%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%B0%D1%8F_%D0%B3%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0&action=edit&redlink=1)). Если же полученное значение U больше табличного, принимается [нулевая гипотеза](https://ru.wikipedia.org/wiki/%D0%9D%D1%83%D0%BB%D0%B5%D0%B2%D0%B0%D1%8F_%D0%B3%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0). Достоверность различий тем выше, чем меньше значение U

5. При справедливости [нулевой гипотезы](https://ru.wikipedia.org/wiki/%D0%9D%D1%83%D0%BB%D0%B5%D0%B2%D0%B0%D1%8F_%D0%B3%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7%D0%B0) (распределение признака в первой выборке соответствует распределению признака во второй выборке) критерий при достаточно большом объёме выборочных данных распределён практически нормально.

Также, можно использовать преобразование данных, например логарифмическое или квадратный корень, чтобы сделать данные более нормально распределенными.

# Глава 3. A/B тестирование, практическая реализация программными средствами

В практической части будут рассмотрены два метода A/B тестирования на одном и том же синтетическом наборе данных. Для реализации будет использован язык программирования Python 3 и некоторые библиотеки для статистического анализа и визуализации данных.

# 3.1. Байесовский метод

Типичная ситуация, в которой оказываются специалисты, состоит в том, что у них есть два варианта рекламы и они хотят выяснить, какой из них более эффективен в долгосрочной перспективе.

Практики называют это A/B-тестированием, а статистики — проверкой гипотез. Рассмотрим следующую проблему. При запуске рекламной онлайн-кампании A в течение определенного периода времени фиксируют положительные результаты, но теперь появилось предположение, что новый вариант рекламы может работать лучше, поэтому необходимо провести эксперимент, разделив аудиторию пополам: одна увидит существующую версию A, а другая новую B. Метрика производительности — конверсия (продажи) за клик. После того, как эксперимент длился два месяца, фиксируются ежедневные клики и конверсии каждой версии необходимо определить какая версия имеет лучшую эффективность.

Моделируем вышеупомянутую проблему с обеими версиями, получающими случайным образом около тысячи кликов в день. Факт, который поможет сгенерировать синтетические данные для этой задачи: гипотетическая кампания B имеет немного лучший коэффициент конверсии, чем A в долгосрочной перспективе (10,5% против 10%).

С помощью этих синтетических данных рассмотрим некоторые полезные статистические концепции и применим их для A/B-тестирования.

Первым этапом работы является подключение необходимых библиотек (рис. 3.1).

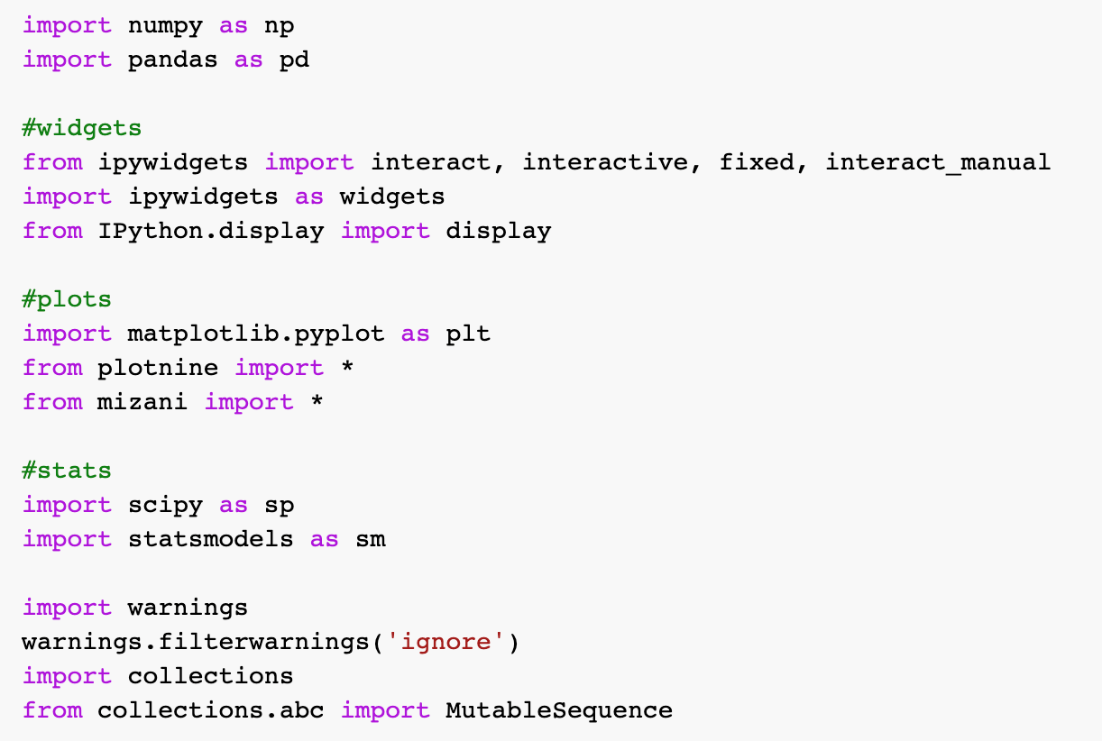
i

Рис.3.1. Пакеты, необходимые для реализации байесовского метода

Функция gen\_campaigns() (рис. 3.2) генерирует фиктивные данные для двух кампаний (A и B) в течение указанного количества дней (заданного параметром nb\_days). Параметры p1 и p2 определяют базовые конверсионные коэффициенты для кампаний A и B соответственно.

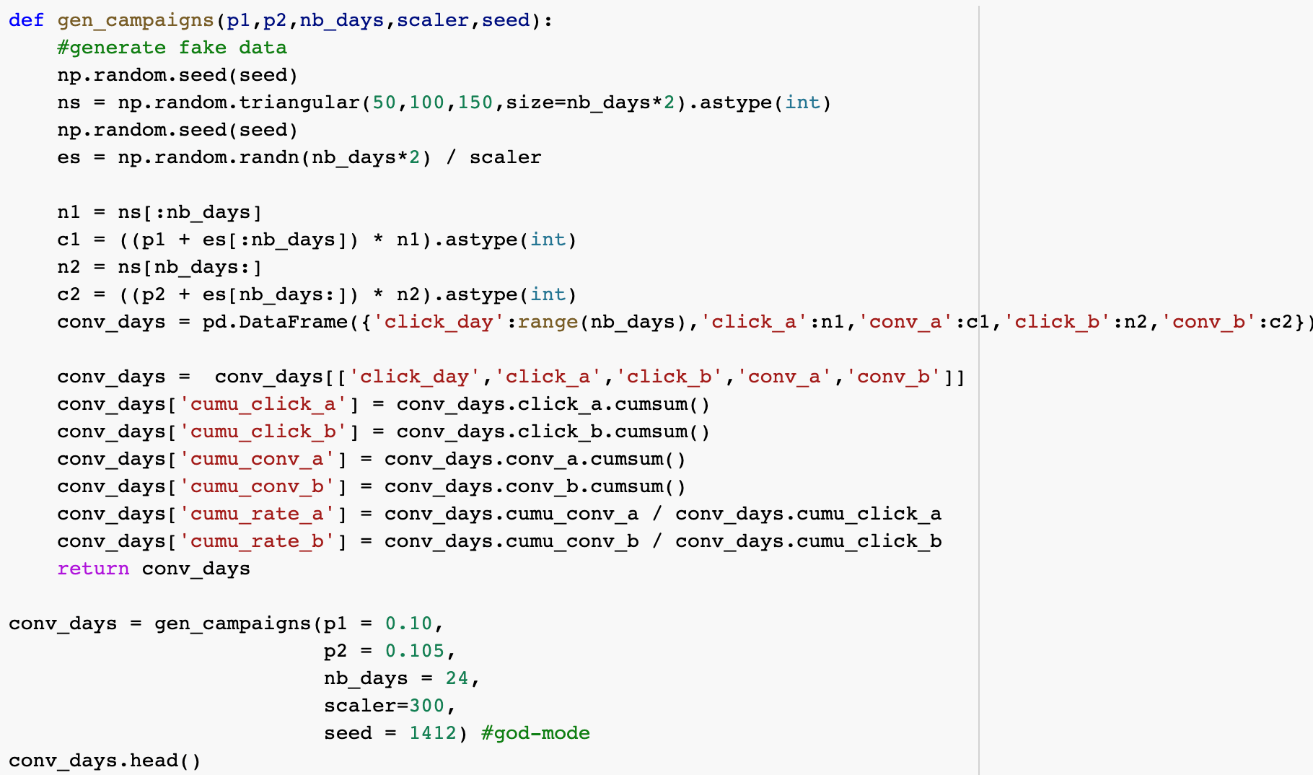


Рис.3.2. Функция для генерации синтетических данных

Затем функция генерирует случайные значения для количества кликов (ns) и случайные значения ошибки (es). Количество кликов ns генерируется с использованием треугольного распределения, а случайные значения ошибки es генерируются с помощью стандартного нормального распределения, которое затем масштабируется заданным параметром scaler.

Далее, функция рассчитывает количество конверсий для каждой кампании (c1 и c2), используя количество кликов и базовые конверсионные коэффициенты для каждой кампании, а также случайные значения ошибки.

Функция создает объект DataFrame (conv\_days), содержащий данные о количестве кликов и конверсий для каждой кампании на протяжении нескольких дней, а также рассчитывает накопленные показатели, такие как кумулятивное количество кликов и конверсий для каждой кампании, а также кумулятивные конверсионные коэффициенты.

Наконец, функция возвращает объект DataFrame conv\_days (рис. 3.3), который затем сохраняется в переменной conv\_days. Функция вызывается с определенными значениями параметров для генерации конкретного набора фиктивных данных.

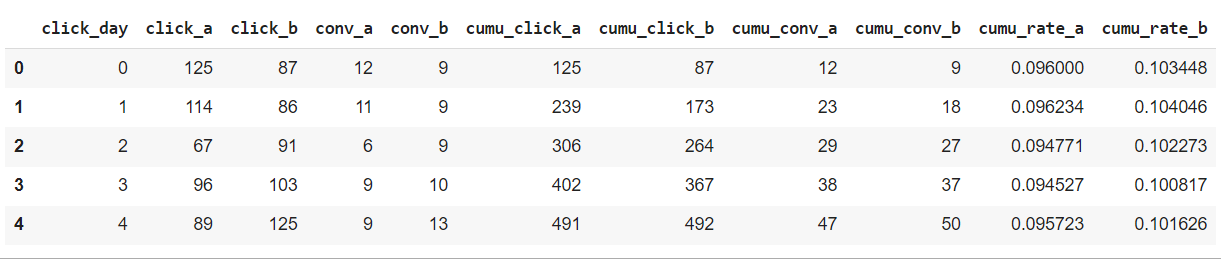


Рис.3.3. Пример сгенерированных данных

Изображение выглядит как диаграмма

Автоматически созданное описание

Рис.3.4. Визуализация конверсионных данных

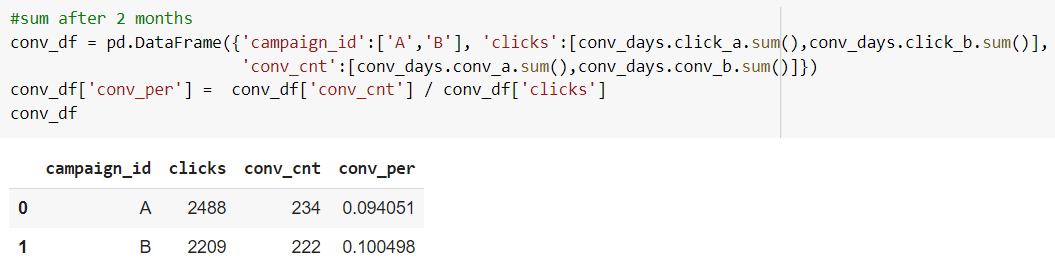


Рис.3.5. Конверсия при длительном эксперименте

P(H|d) Апостериорная вероятность. Это вероятность того, что какая-то конкретная гипотеза верна после сбора и добавления новых данных. Рассуждая в терминах А\В тестирования, она выражает вероятность получить весь набор возможных исходов теста. Например, коэффициент конверсии для A лучше, чем для B, на +0,1%, +0,2%, -0,1%, -0,2%, … и каждый промежуточный пункт и далее. Множество всех возможных исходов бесконечно, а это значит, что вероятностное поле исходов будет представлено непрерывным распределением.

P(d|H) функция правдоподобия. Она показывает, насколько вероятно наблюдать собранные данные для заданной гипотезе модели, представлена непрерывным распределением

P(H) априорное распределение. Оно кодирует информацию о модели до появления реальных данных. Например, предварительная информация о том, что средний коэффициент конверсии для A должен быть сосредоточен вокруг 10% со стандартным отклонением в 1%. Тогда предварительная оценка для A будет представлена соответствующим нормальным распределением. Поскольку нет информации о B, лучше всего предположить, что B ведет себя так же, как A.

Рассмотрим подробнее функцию правдоподобия – это вероятность того, что специалист получит этот набор данных о кликах и конверсиях при условии, что истинный коэффициент конверсии равен p, её можно описать как функцию массы вероятности распределения Бернулли:

где – бинарный флаг конверсии, P – вероятность гипотезы H:

где 𝛼 и 𝛽 являются гиперпараметрами, коррелирующими с количеством успехов и неудач соответственно, а 𝐵(𝛼,𝛽) это бета-функция нормализованная между 0 и 1.

Интуитивно бета-распределение формируется количеством успехов и неудач, инициализированных 𝛼 и 𝛽; более того, более высокое число означает большую уверенность в распределении, что приводит к меньшей дисперсии (рис. 3.6).

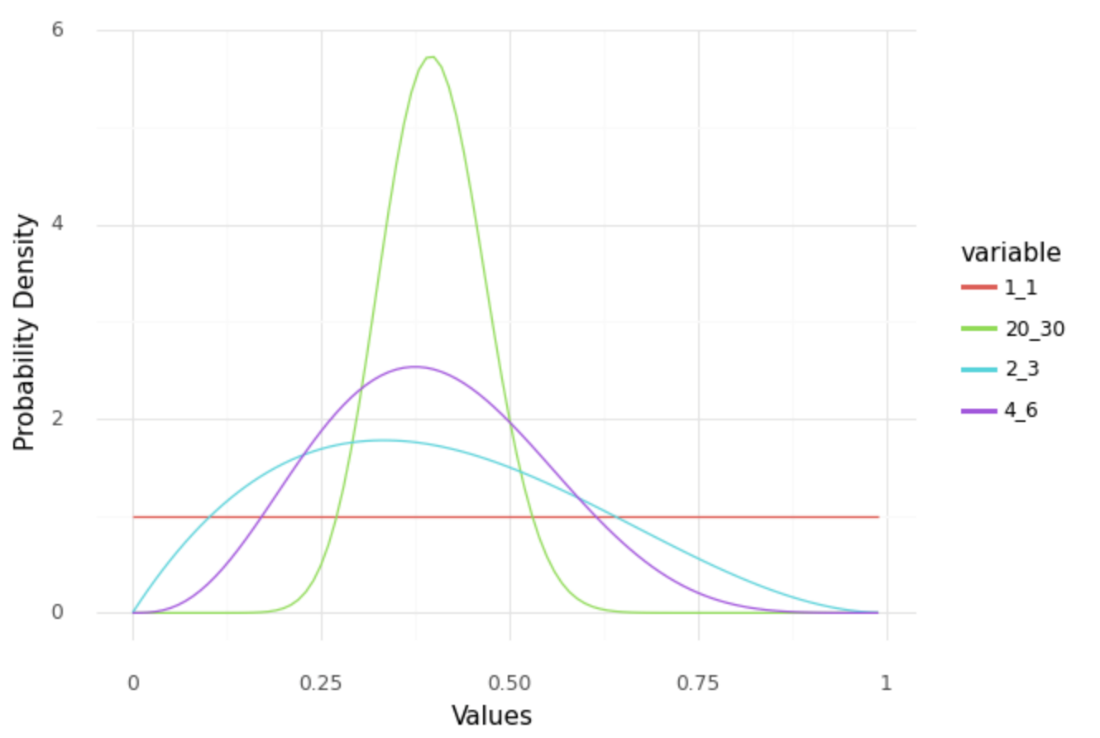


Рис.3.6. Плотность вероятности при различных значениях коэффициентов

Важно отметить, что 𝑃(𝐷) и нормирующий множитель 𝐵(𝛼,𝛽) являются константами в отношении гипотезы, поэтому верным решением будет обойти их и считать, что апостериорное распределение пропорционально:

Результат похож на не нормализованное бета-распределение

Для событий, которые ведут себя как испытания Бернулли, апостериорное распределение принимает очень удобную форму. Все, что нужно сделать - выбрать разумный набор 𝛼 и 𝛽 для априорной вероятности. Именно здесь предположение влияет на байесовское A/B тестирование. Один из способов - установить априорную вероятность на основе результатов прошлых результатов; например, если подобный эксперимент уже был проведён и средний коэффициент конверсии составил 10%, следует изменить масштаб априорного распределения в зависимости от того, насколько коэффициент конверсии можно считать достоверным. Разные распределения могут представлять 10%, например 𝐵𝑒𝑡𝑎(1,9) и 𝐵𝑒𝑡𝑎(10,90) (рис. 3.7) , но с разным уровнем уверенности. 𝐵𝑒𝑡𝑎(10,90) является более острым около 10%, чем 𝐵𝑒𝑡𝑎(1,9) .

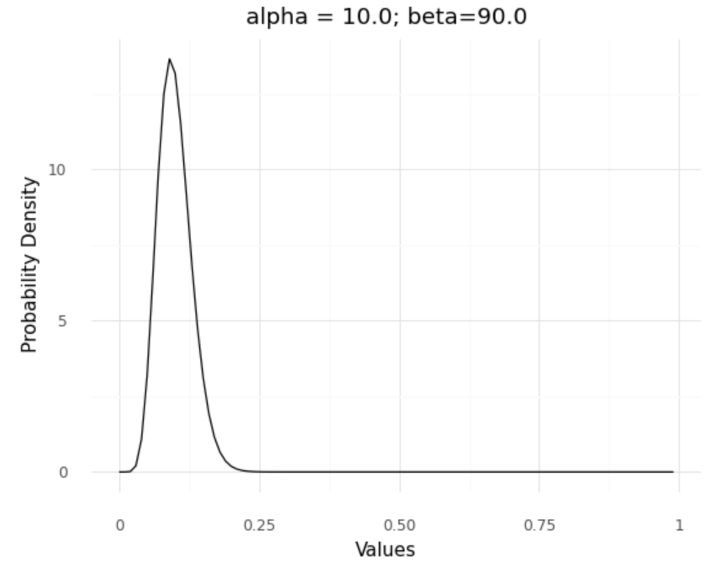
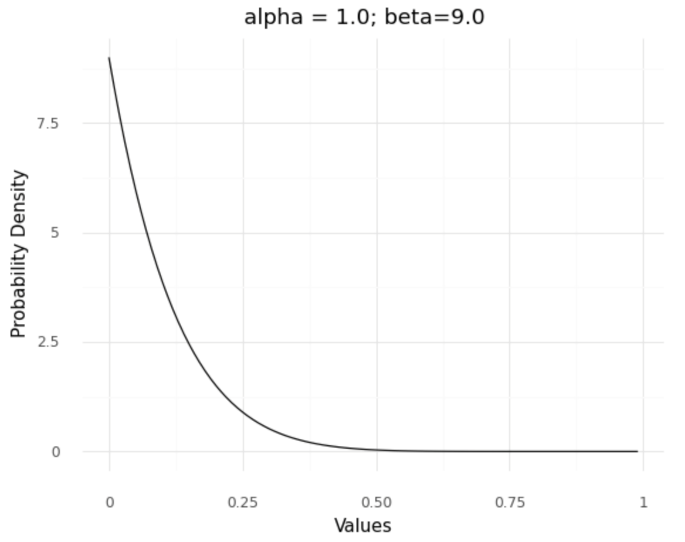


Рис.3.7. Плотность вероятности при различных значениях коэффициентов

В зависимости от априорных значений для A и B, по-разному формируются их апостериорные распределения. Априорное значение влияет на достоверность значений апостериорных значений (точечность) и на ожидаемые значения самих апостериорных значений. Например, даже если наблюдаемый коэффициент конверсии составляет около 10%, если дать абсурдно сильную априорную оценку, такую как 𝐵𝑒𝑡𝑎(999,1) , что означает коэффициент конверсии 99,9 %, апостериорные ожидания могут составлять около 30 %.

Функция sample\_proportion() (рис.3.8) генерирует случайную выборку из бета-распределения, которое моделирует вероятность конверсии для версии A или B.

Функция proportion\_test\_b() использует sample\_proportion() для генерации выборок конверсий для версии A и B и вычисляет вероятность того, что конверсии версии A больше, чем у версии B. Эта вероятность является мерой статистической значимости различия между двумя версиями рекламы.

Функция proportion\_ratio() использует sample\_proportion() для генерации выборок конверсий для версии A и B и вычисляет отношение конверсии A к конверсии B.

Функция proportion\_ci\_b() использует proportion\_ratio() для вычисления доверительного интервала для отношения конверсий A/B с использованием заданного уровня значимости и гиперпараметров бета-распределения.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рис. 3.8. Функция sample\_proportion() для генерации случайной выборки из бета распределения

Примерно для 80% отношения распределения A хуже, чем B, и в случае ошибки, A не более чем на 20% лучше, чем B. И аналогично частотному подходу, можно вычислить диапазон, в который попадает большинство этих отношений (предположим, 95%), называется доверительным интервалом (рис.3.9).

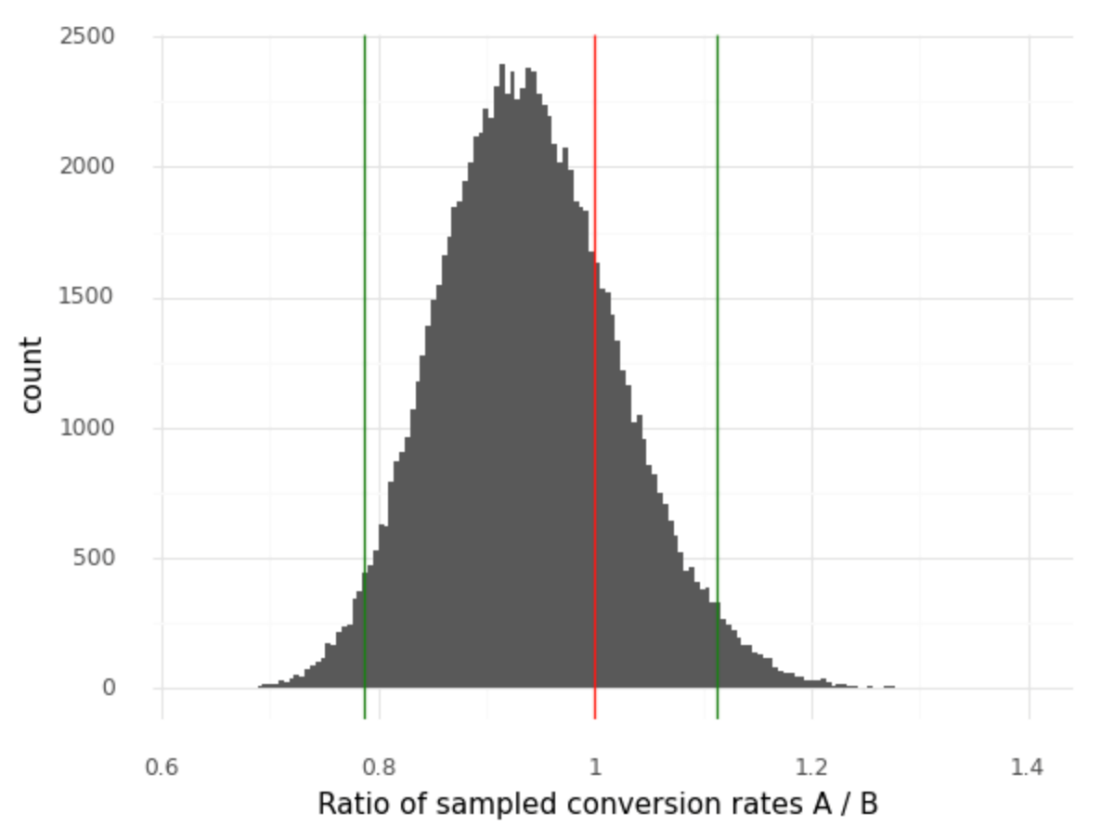


Рис.3.9 Визуализация доверительного интервала.

Свобода от частотных p-значений позволяет не сталкиваться с тем, что бесконечное количество выборок всегда дает статистическую значимость, даже если истинные значения точно такие же. Как видно из приведенного ниже графика, вероятность того, что A превзойдет B в результате выборки методом Монте-Карло из апостериорных данных, остается около 50%, когда нет истинной разницы, и постепенно уменьшается, когда она есть.

Сгенерируем две рекламные кампании (conv\_days2 и conv\_days3), каждая из которых продолжается в течение 60 дней. conv\_days2 и conv\_days3 имеют одинаковый уровень конверсии для A, но разный уровень конверсии для B. Затем используем функцию proportion\_test\_b, чтобы вычислить вероятность того, что A лучше, чем B, для каждого дня в каждой из этих двух кампаний. Результаты сохраняются в столбцах 'prob\_same' и 'prob\_diff' для conv\_days2 и conv\_days3 соответственно.

Сразу же возникает вопрос: когда нужно остановиться и объявить победителя. Технически, когда есть истинное различие, нужно остановиться в любой момент времени, и в конечном итоге определить победителя, но с разной степенью вероятности. Можно использовать простое правило, например, если вероятность того, что A победит B, ниже 10% или выше 90%, нужно объявить победителя, как показано на графике выше.

Критерий остановки для таких экспериментов – ROPE; рассмотрим понятие оставшейся стоимости, введенное Google. Оставшаяся стоимость за раунд эксперимента определяется как:

По мере продолжения эксперимента строится график распределения 𝑉𝑡 и останавливается, когда 1-𝛼 процентиль ниже порога. Интуитивно это означает, что на 1-𝛼 % уверенность в том, что "лучшая" версия может быть побеждена с отрывом, равным пороговому значению. Для практических целей используют 95-й процентиль и порог в 1 %.

Иными словами, необходимо найти «оставшуюся стоимость» (value remaining) для двух рекламных кампаний A и B, которые представляют собой долю преобразований (conv\_cnt) в общем числе кликов (clicks).

Функция (рис.3.10) сначала использует функцию sample\_proportion, чтобы смоделировать распределение вероятностей для доли преобразований A и B. Затем она вычисляет максимальную долю преобразований, выбирает лучшую кампанию (A или B), и вычисляет её долю преобразований.



Рис.3.10. Функция вычисления оставшейся стоимости

Затем функция вычисляет оставшуюся стоимость, определяемую как разница между максимальной долей преобразований и долей преобразований лучшей кампании, деленная на долю преобразований лучшей кампании. Эта величина отражает насколько хуже результаты лучшей кампании, чем бы они могли быть при использовании наилучшей стратегии.

Наконец, функция возвращает значение оставшейся стоимости в заданном квантиле q (по умолчанию 95%).

Нужно заметить, что в случае, когда истинная разница составляет 1%, оставшееся значение постепенно уменьшается (рис.3.11) ниже порога в 1%. С другой стороны, когда истинная разница составляет 0%, оставшееся значение всегда колеблется около 10%.

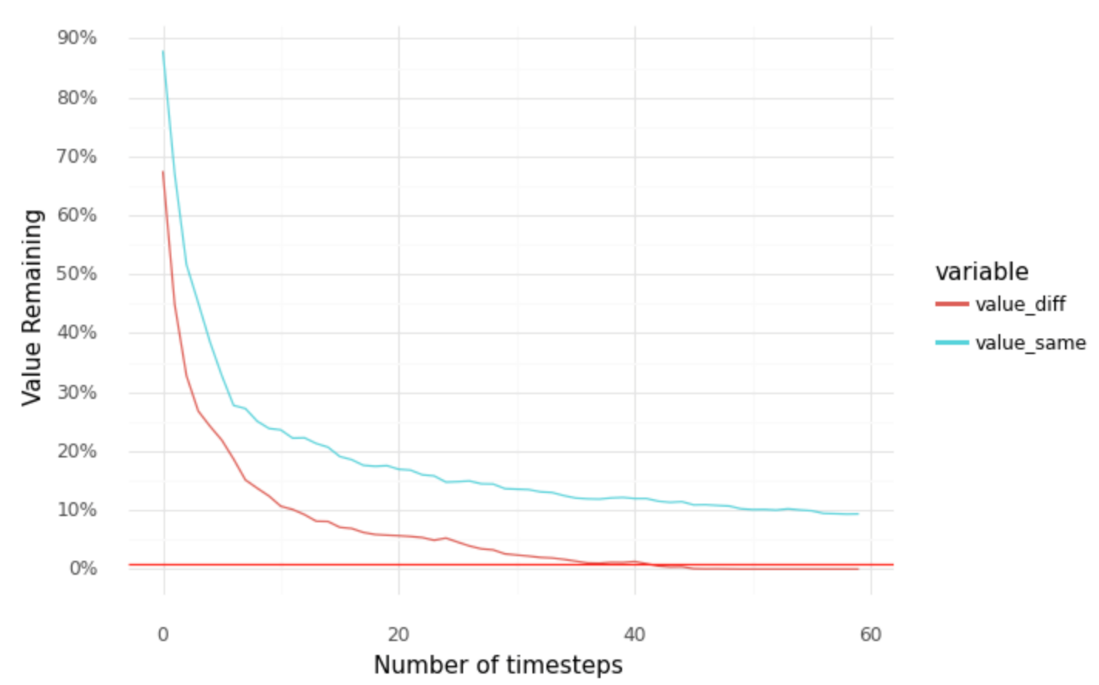


Рис.3.11. Визуализация разницы

# 3.2. Частотный метод

Рассмотрим частотный метод на тех же данных, используем предположение о том, что конверсия в версии В лучше и составляет 10,5% (в версии А 10%), для этого загрузим пакеты (Рис 3.12):

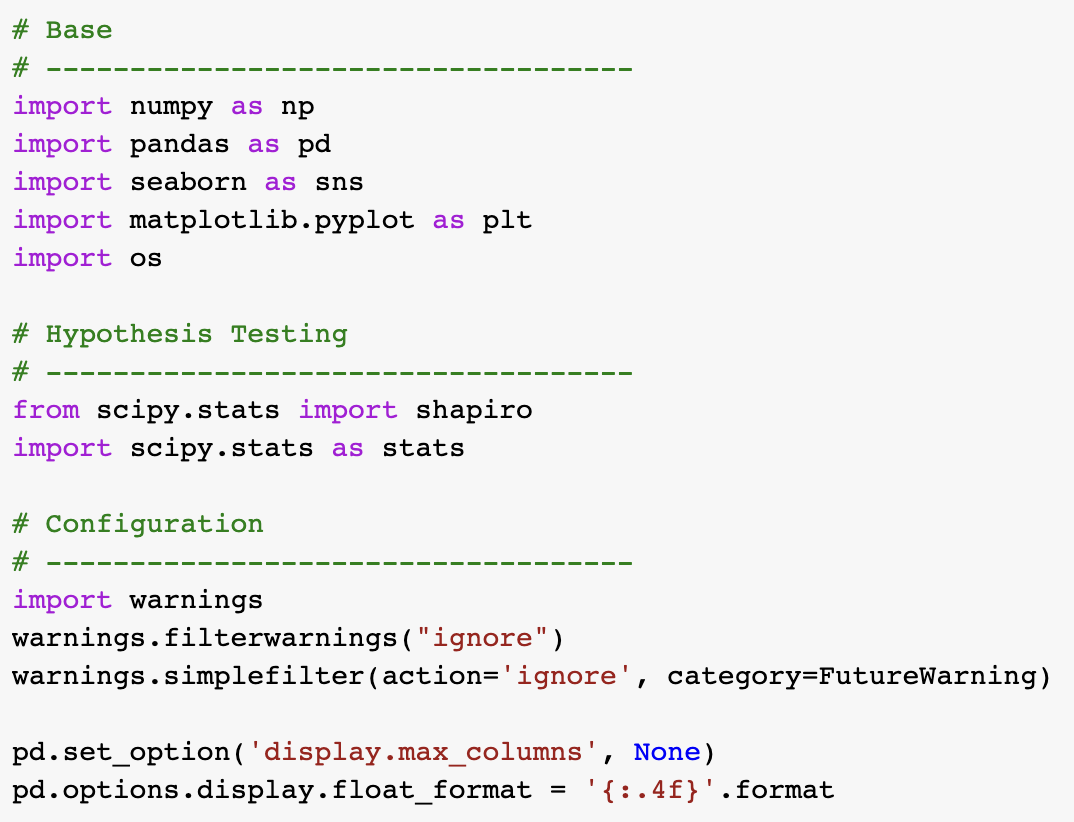


Рис.3.12. Пакеты, необходимые для реализации частотного метода

Сгенерируем данные (рис 3.13).

Изменим формат представления данных для более удобного анализа в дальнейшем, создадим таблицу, в которой будет храниться количество кликов, конверсии и метка версии (рис.3.14).

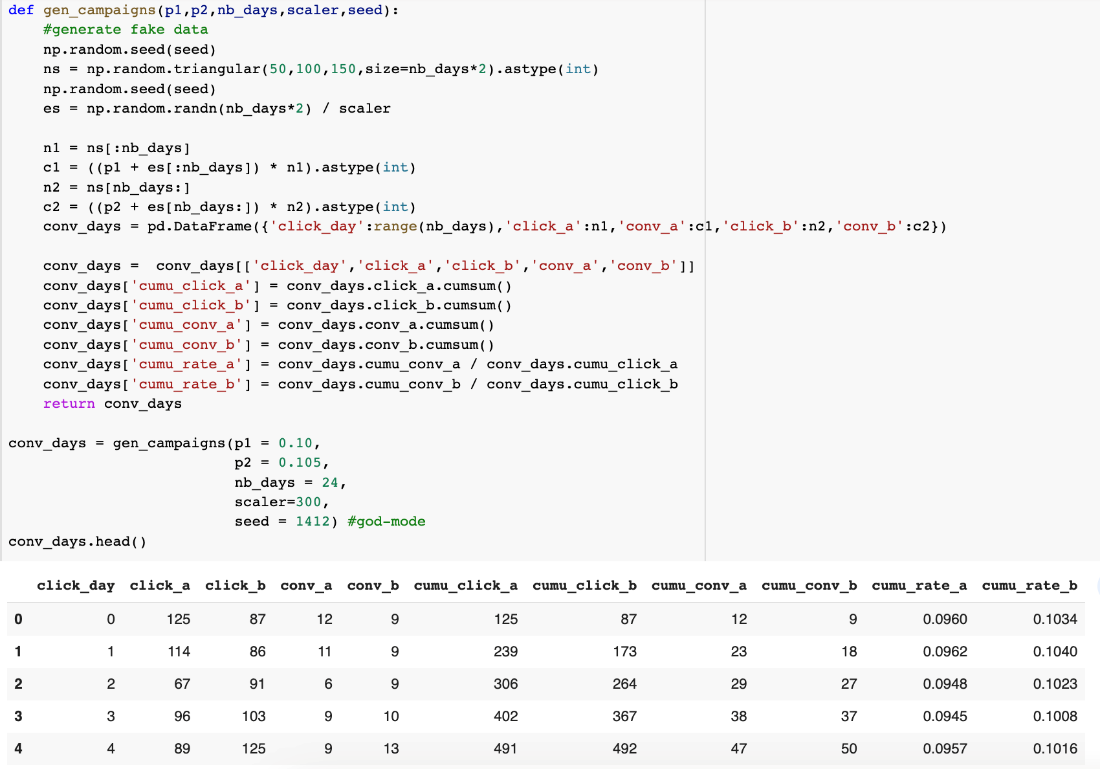


Рис.3.13. Функция для генерации синтетических данных

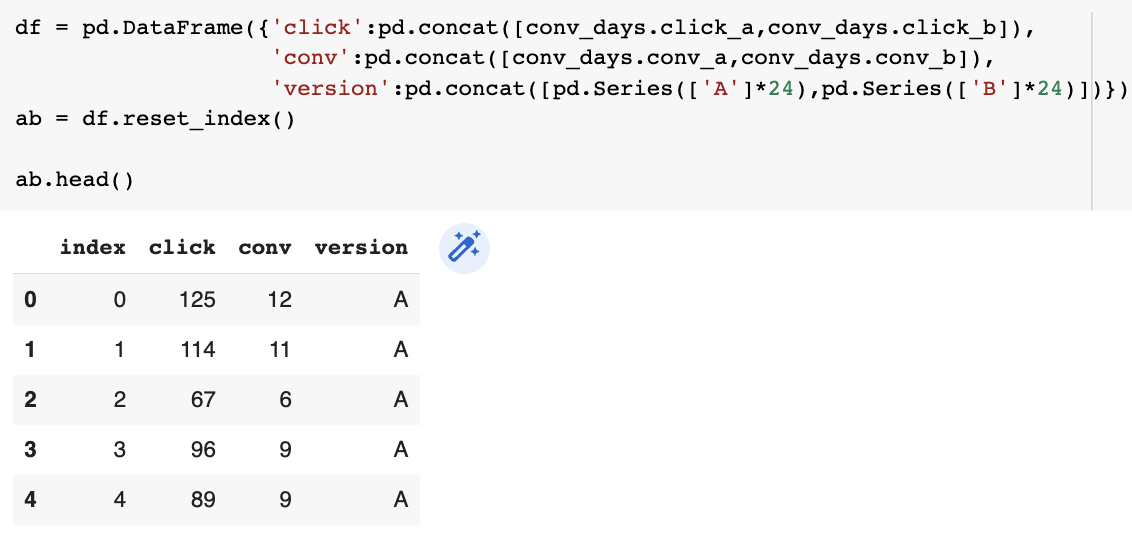


Рис. 3.14. Таблица с данными

Следует рассмотреть подробнее статистику (рис 3.15 и рис 3.16).

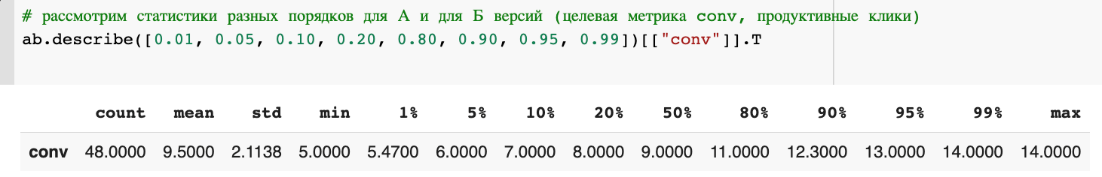


Рис. 3.15. Демонстрация статистик

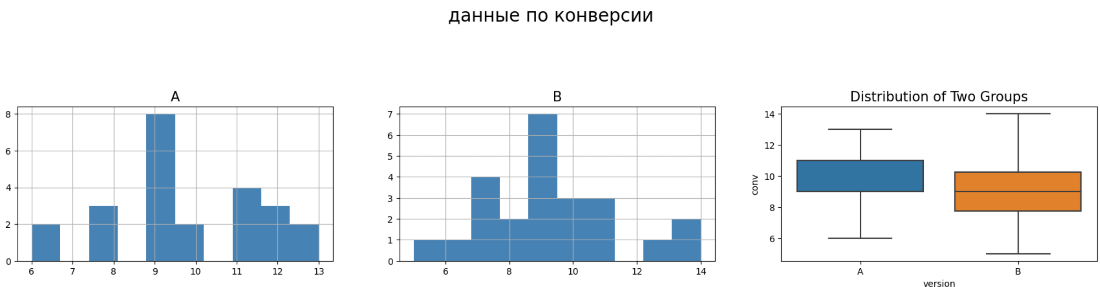


Рис. 3.16. Визуализация групп данных

В графе box создается коробка от первого квартиля до третьего квартиля. Здесь также есть вертикальная линия, которая проходит через коробку на медиане. Здесь ось x обозначает данные, а ось y показывает частотное распределение. boxes: основная часть boxplot, показывающая квартили и доверительные интервалы медианы, если они включены. медианы: горизонтальные линии в медиане каждого прямоугольника. усы: вертикальные линии, простирающиеся до самых крайних, не выделяющихся точек данных. caps: горизонтальные линии на концах усов. fliers: точки, представляющие данные, которые выходят за пределы усов (fliers).

Проведём предварительные тесты на нормальность распределения (рис. 3.17).

Далее используем функцию для автоматического анализа статистической значимости (Рис 3.18, 3.19, 3.20)

Эта функция принимает фрейм данных, содержащий данные для двух групп A и B, а также имена столбцов, содержащих результаты теста, и возвращает результаты статистического анализа A/B-тестирования. Функция сначала разделяет группы A и B на отдельные серии данных, а затем выполняет проверку нормальности распределения для каждой из серий данных с помощью теста Шапиро-Уилка. Если распределения нормальны, функция выполняет параметрический тест Стьюдента для независимых выборок с предположением о равенстве дисперсий между группами (если тест Левена не отвергает нулевую гипотезу о равенстве дисперсий) или без предположения о равенстве дисперсий (если тест Левена отвергает нулевую гипотезу о равенстве дисперсий).

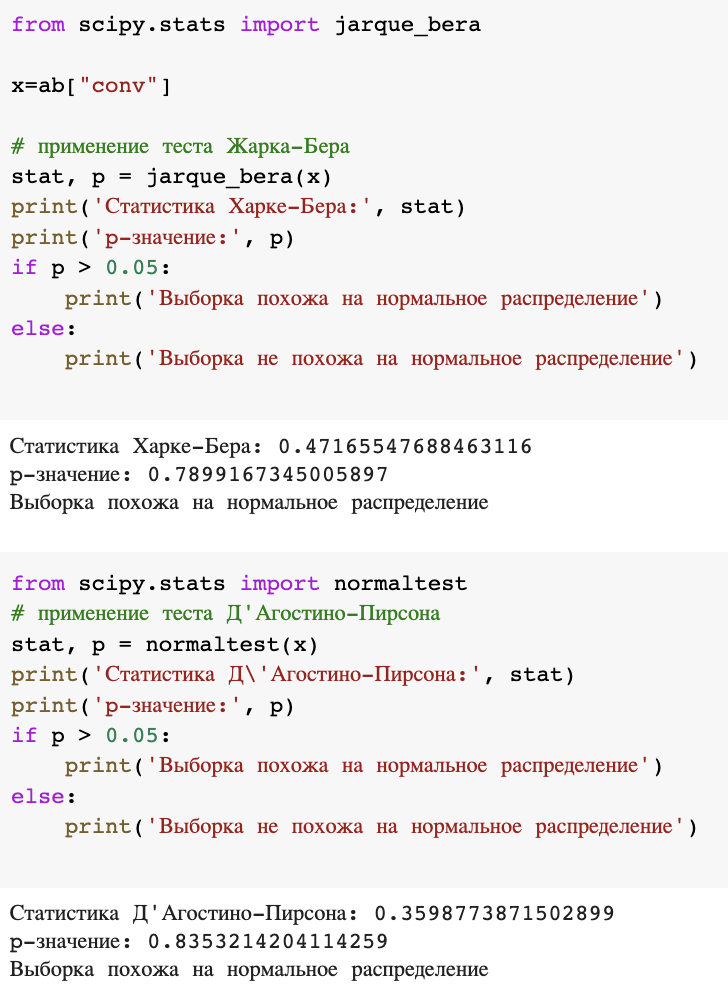


Рис. 3.17. Тесты на нормальность распределения

Если распределения не являются нормальными, функция выполняет непараметрический тест Манна-Уитни. Функция возвращает результаты теста, включая значение p-уровня значимости и вывод, позволяющий определить, отвергается ли нулевая гипотеза. Результаты также включают комментарий о том, являются ли группы A и B схожими или нет.

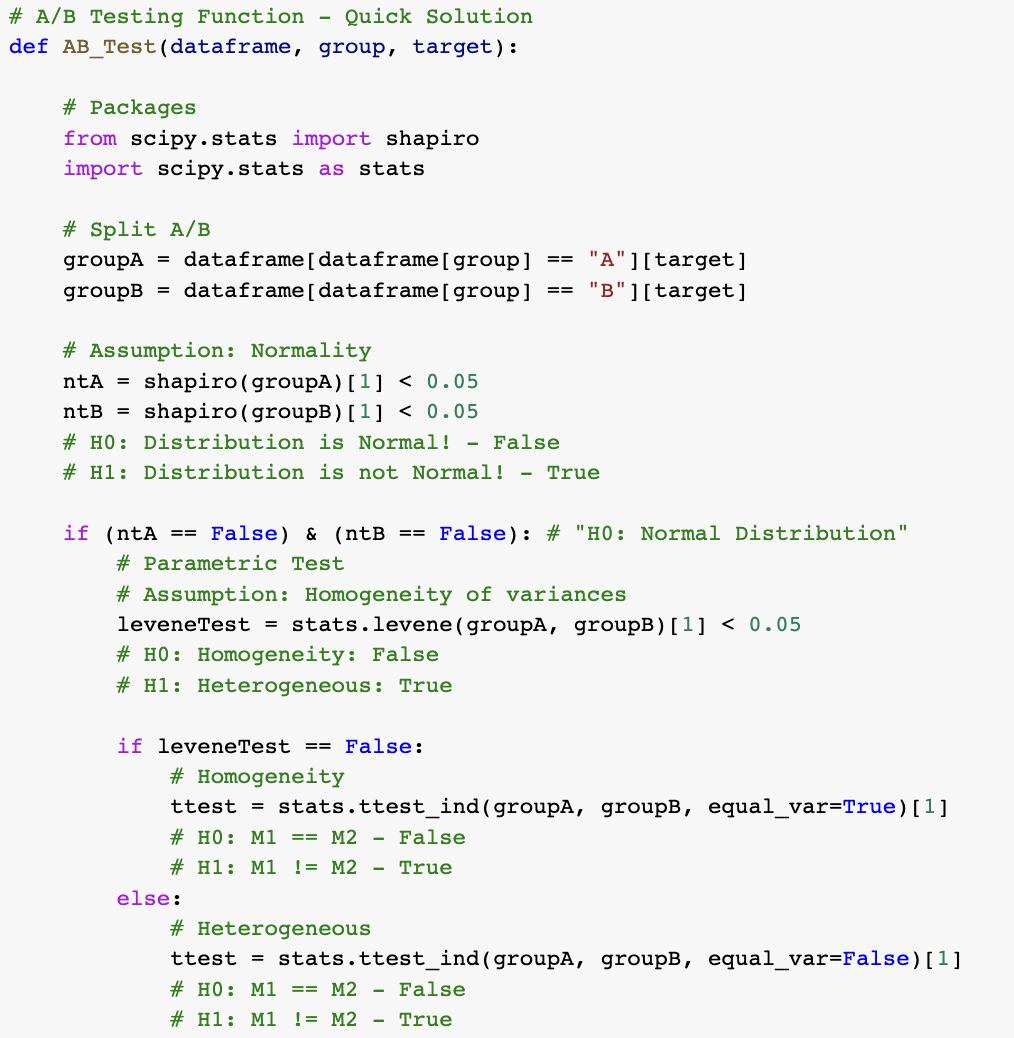


Рис. 3.18. Функция для автоматического анализа статистической значимости



Рис. 3.19. Функция для автоматического анализа статистической значимости(продолжение)

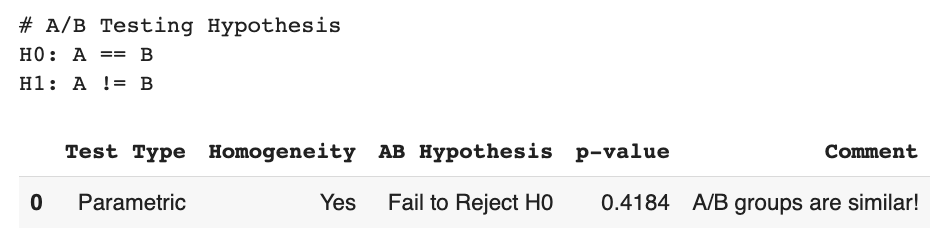
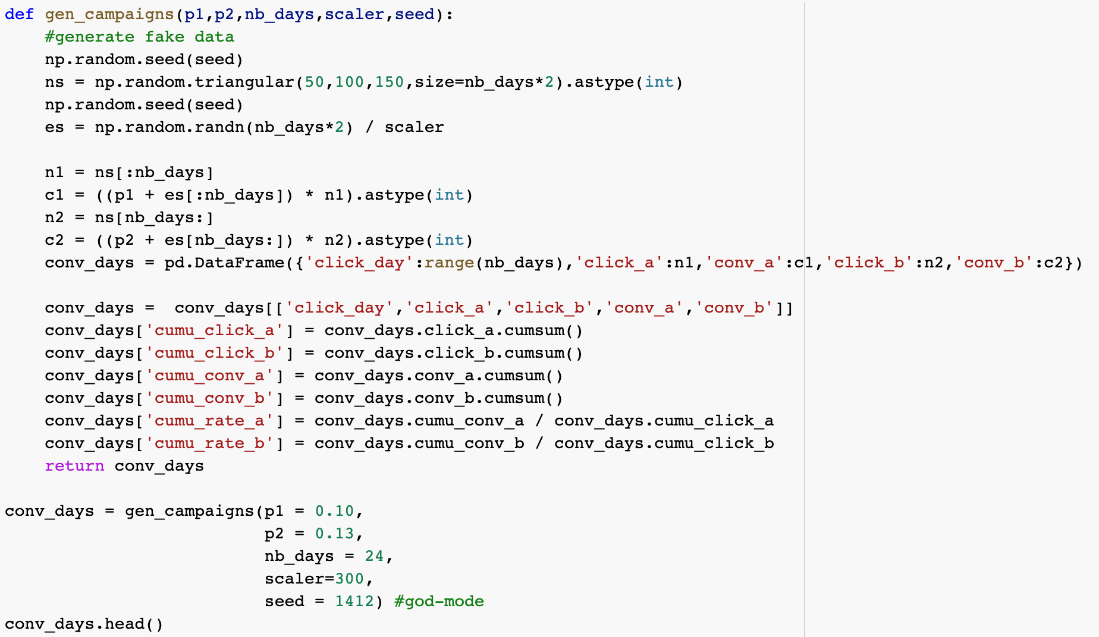


Рис. 3.20. Результаты проверки

Тест не видит статистически значимой разницы между выборками, дело в изначальной конфигурации данных конверсия в версии В лучше и составляет 10,5% (в версии А 10%)

Теперь проверим работу теста с немного большей разницей в проценте конверсии, теперь не 10% и 10,5%, а 10% и 13% (рис 3.21).



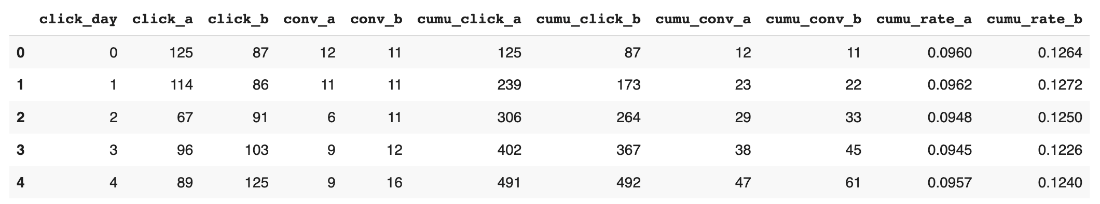


Рис. 3.21. Функция для генерации синтетических данных

Следует рассмотреть подробнее статистику (рис 3.22 и рис 3.23).

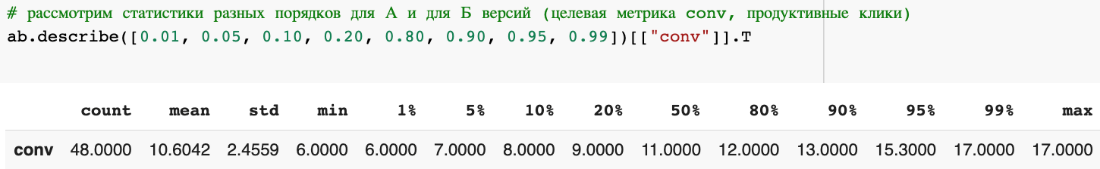


Рис. 3.22. Демонстрация статистик

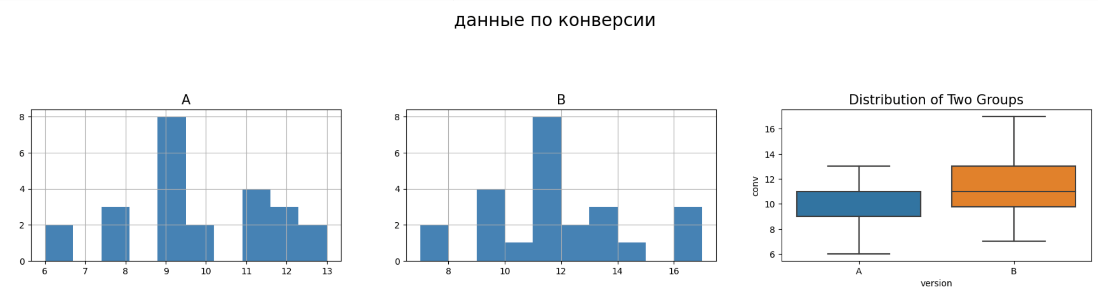


Рис. 3.23. Визуализация групп данных

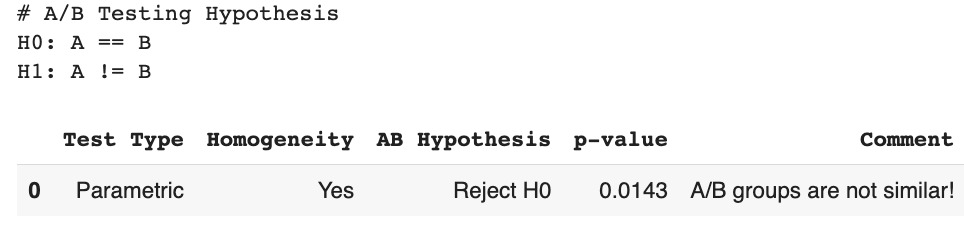


Рис. 3.24. Результаты теста

При новой конфигурации данных тест находит статистически значимую разницу. Что подтверждает меньшую (в сравнении с байесовским подходом) чувствительность частотного подхода.

В реальной практике исследователи не всегда пользуются статистическими тестами, в зависимости от специфики задачи могут быть использованы различные методы оценки значимости различий между выборками. Далее будет более подробно рассмотрен один из них – бутстрэп.

В рамках A/B тестирования обычно проверяют гипотезу о равенстве квантилей на уровне значимости 5%, достаточно построить 95% доверительный интервал для разности квантилей между группами (рис 3.25, 3.26, 3.27). Если ноль находится вне доверительного интервала, то отличия статистически значимы, иначе нет.

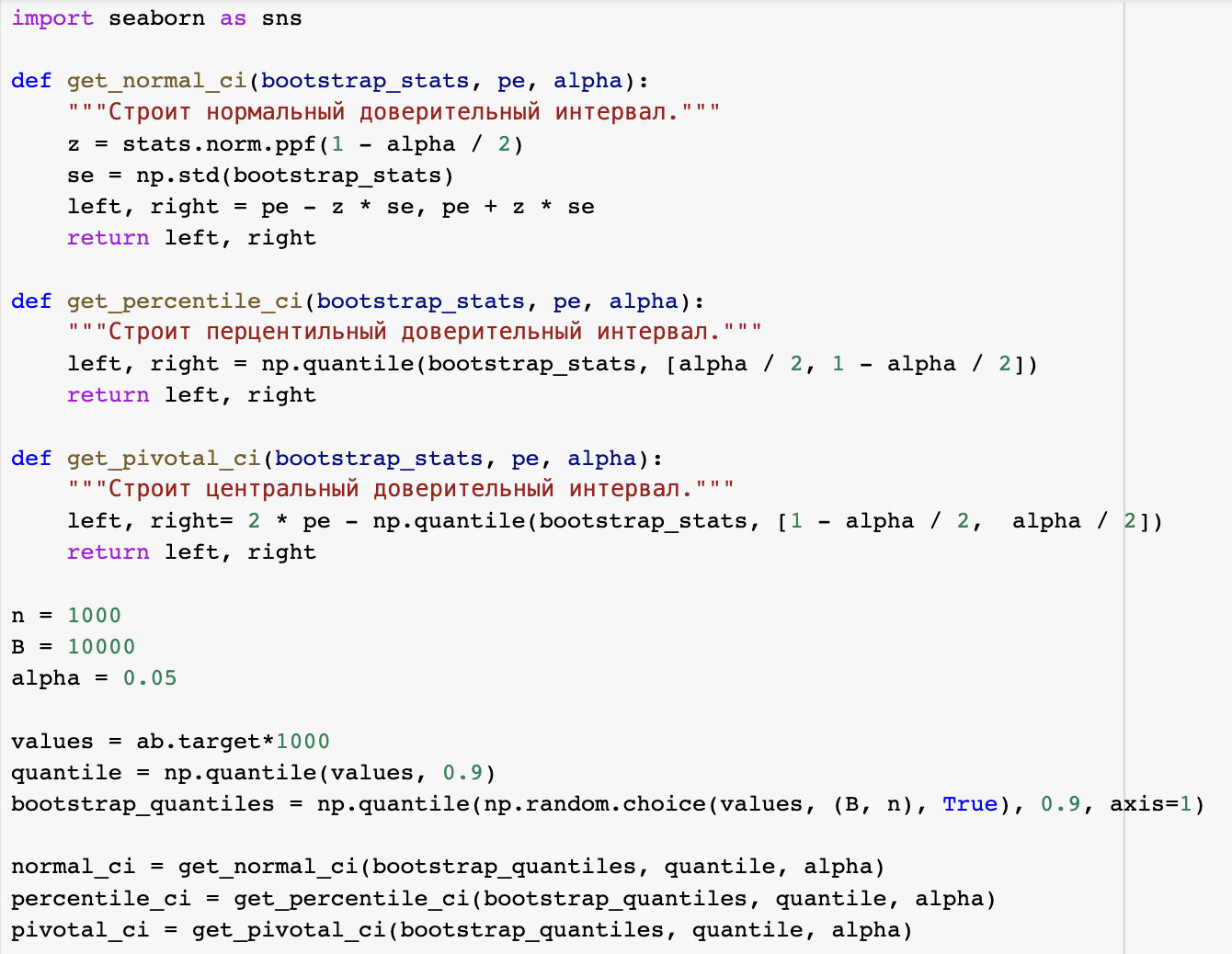


Рис. 3.25. Функции для реализации и визуализации доверительных интервалов

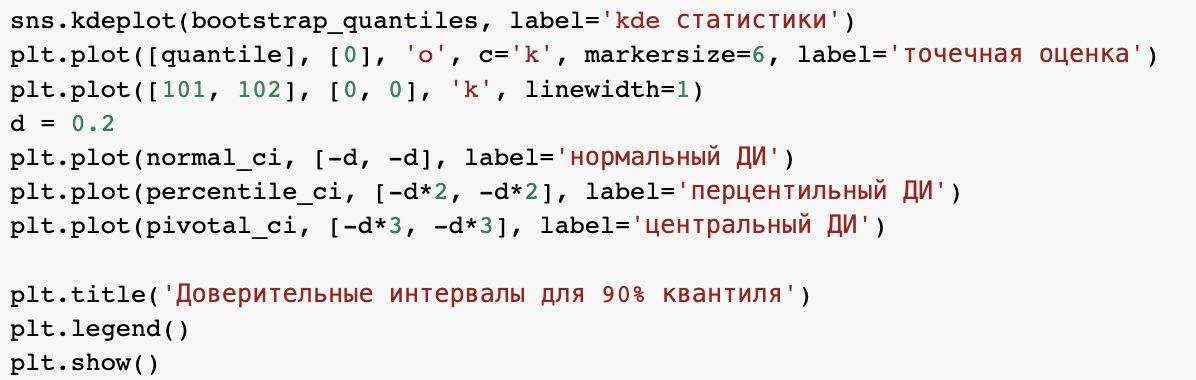


Рис. 3.26. Функции для реализации и визуализации доверительных интервалов(продолжение)

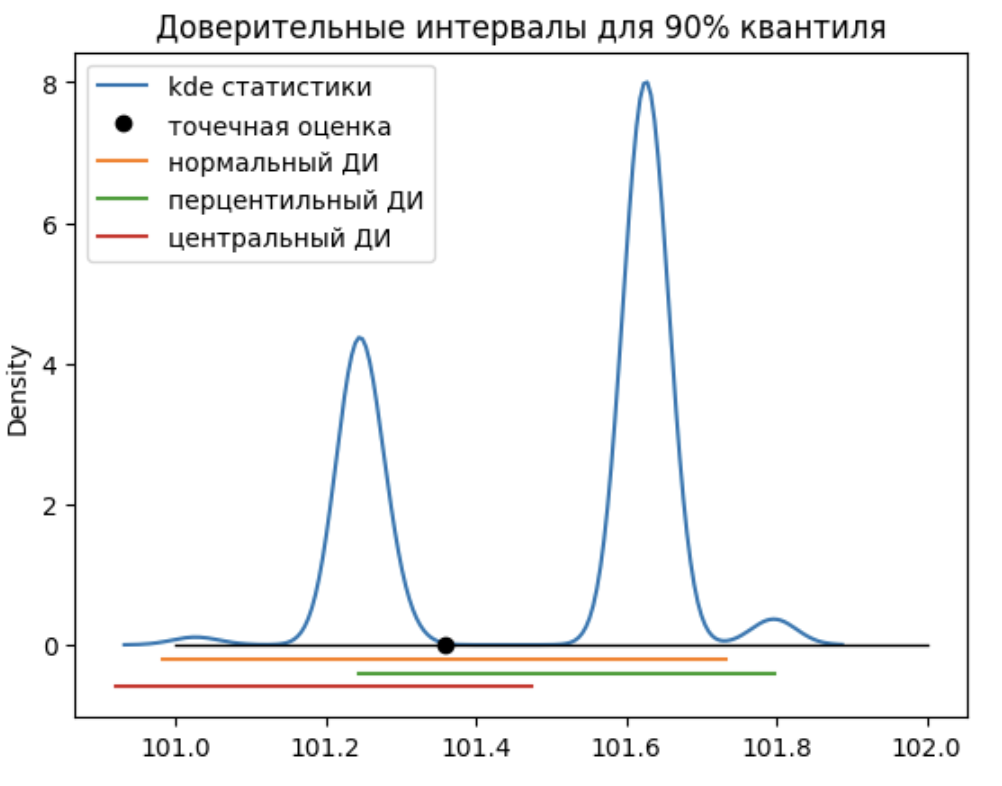


Рис. 3.27. Визуализация доверительных интервалов

Можно сделать вывод о том, что бутстрэп более чувствителен к различиям в выборке, чем статистические тесты (рис 3.28), тем не менее в виду вычислительной сложности его можно эффективно применять для выборок, относительно не больших размеров.

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рис. 3.28. Бутстрэп для A/B теста

# 3.3. А/В тестирование: реальный данные

Как было показано ранее, статистические тесты, лежащие в основе частотного подхода, могут быть не слишком чувствительны к выборкам с небольшой разницей, в коммерческой практике встречаются различные случаи, эксперты рекомендуют использовать частотный подход, если гипотеза об изменении целевой метрики достаточно сильная. Исследуя реальные данные, специалисты часто сталкиваются с длительной очисткой и предобработкой данных, эти процессы на примере реальных данных из мобильного приложения будут в дальнейшем продемонстрированы на практике.

Cookie Cats - популярная мобильная игра-головоломка, разработанная компанией Tactile Entertainment. Это классическая головоломка в стиле «соедини три», в которой игрок должен соединить плитки одного цвета, чтобы очистить поле и выиграть уровень.

По мере прохождения игры игроки будут сталкиваться с рекламными банерами, которые заставят их подождать некоторое время, прежде чем они смогут продвинуться дальше или совершить покупку в приложении. В этом проекте проанализированы результаты A/B-теста, в ходе которого первый рекламный банер в Cookie Cats были перенесены с 30-го на 40-й уровень. В частности, выполняется анализ влияния на удержание игроков и количество пройденных раундов игры.

Данные получены от 90 189 игроков, которые установили игру во время проведения AB-теста. Данные содержат следующую информацию:

* userid - уникальный номер пользователя;
* version - определяет в контрольной(gate\_30) или тестовой(gate\_40) группе находиться человек;
* sum\_gamerounds - количество раундов, проигранных игроком;
* retention\_1 - пользователь вернулся спустя 1 день после установки;
* retention\_7 - пользователь вернулся спустя 7 дней после установки.

Первым этапом работы является подключение необходимых библиотек и загрузка данных (рис. 3.29-3.32).

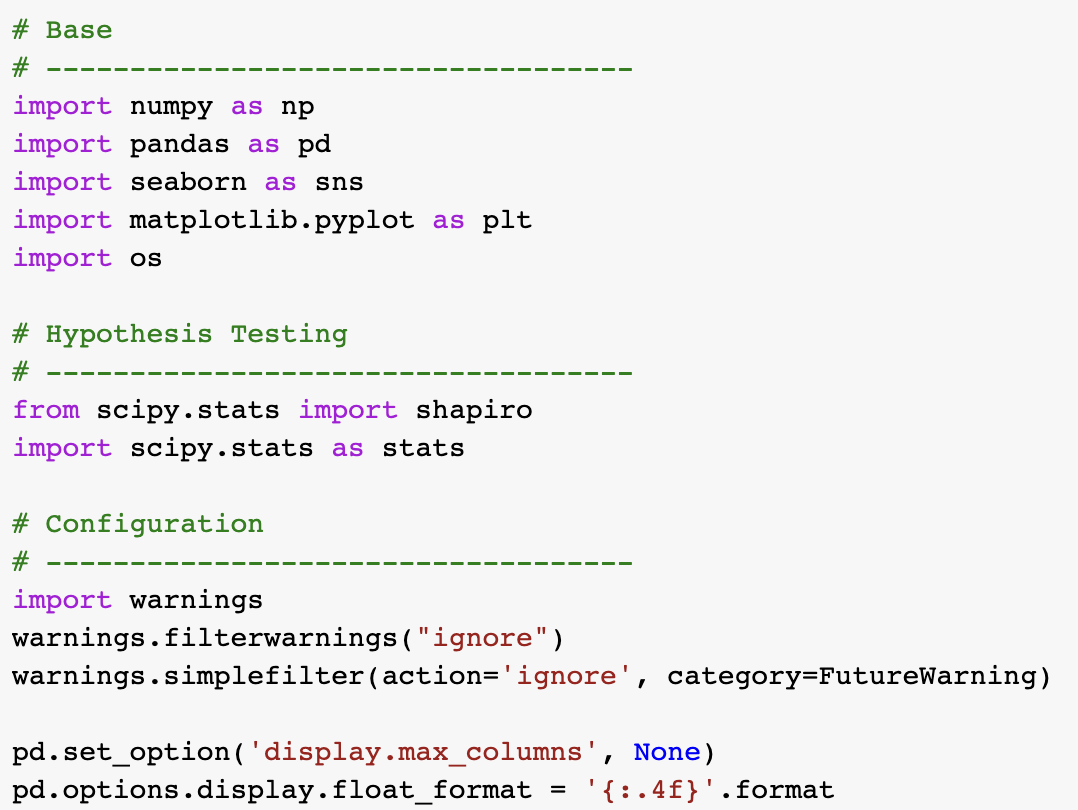


Рис. 3.29. Программные пакеты для частотного подхода



Рис. 3.30. Функция для корректной загрузки данных из CSV файла



Рис. 3.31. Функция для корректной загрузки данных из CSV файла (продолжение)

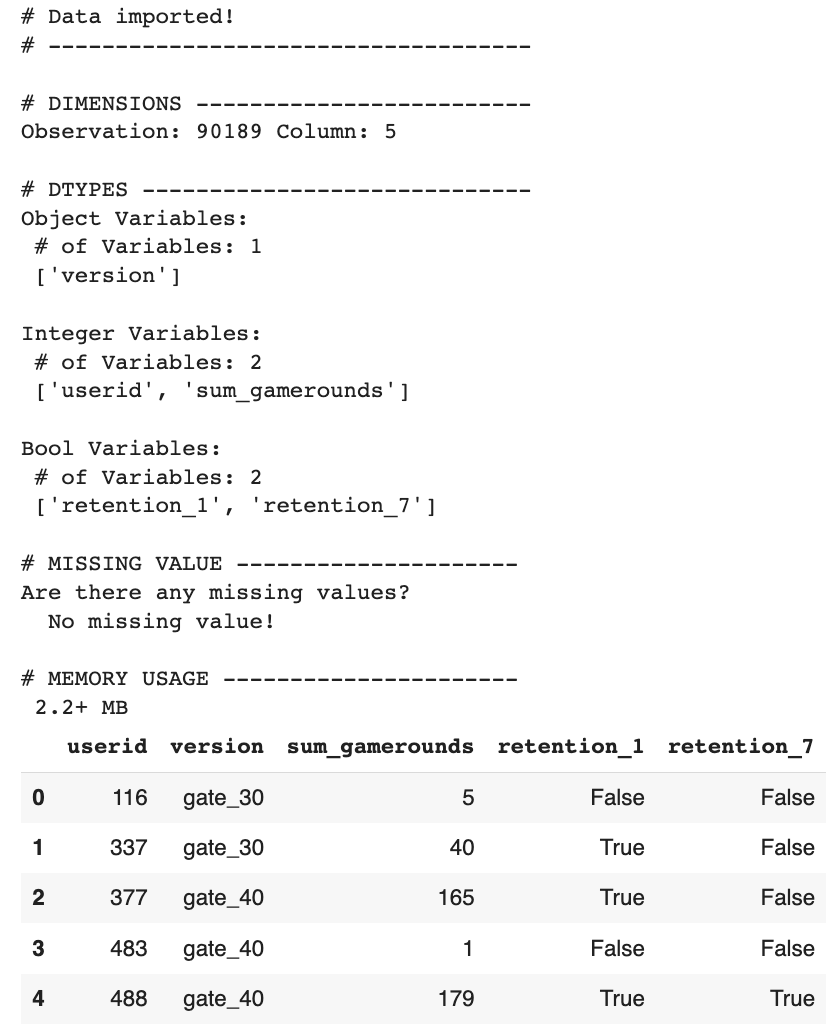


Рис. 3.32. Результат загрузки

Удаляем все строки с пропусками (NAN) в данных и проверяем уникальность ID пользователей, результат проверки - True все идентификаторы уникальны (рис 3.33).

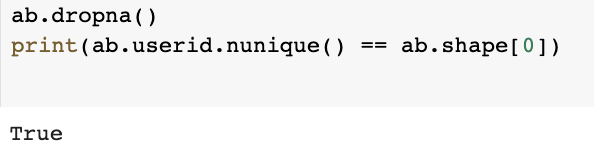


Рис. 3.33. Удаление строк с пропусками

Следующий этап оценка статистик и визуализация (рис. 3.34, 3.35).

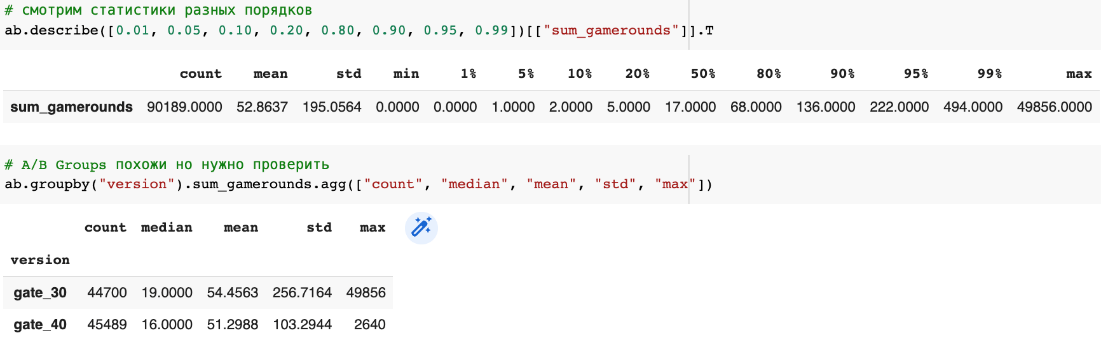


Рис. 3.34. Демонстрация статистик, как всех данных, так и по группам

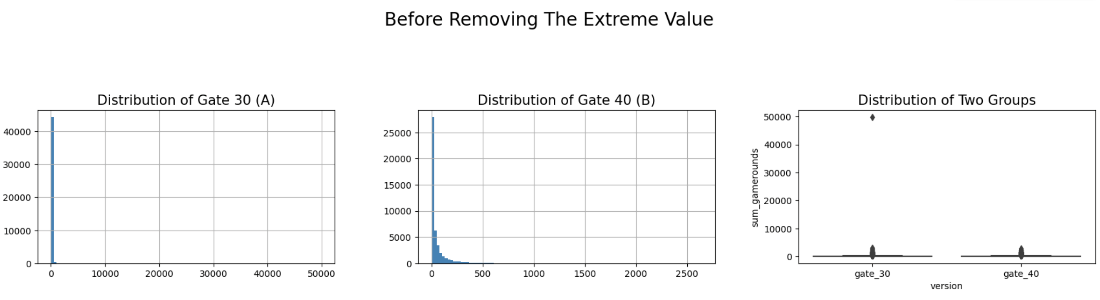


Рис. 3.35. Визуализация данных

Предполагаем, что данные распределены не нормально, в одной из групп явно выделяется выброс (рис 3.36), далее рассмотрим подробнее.

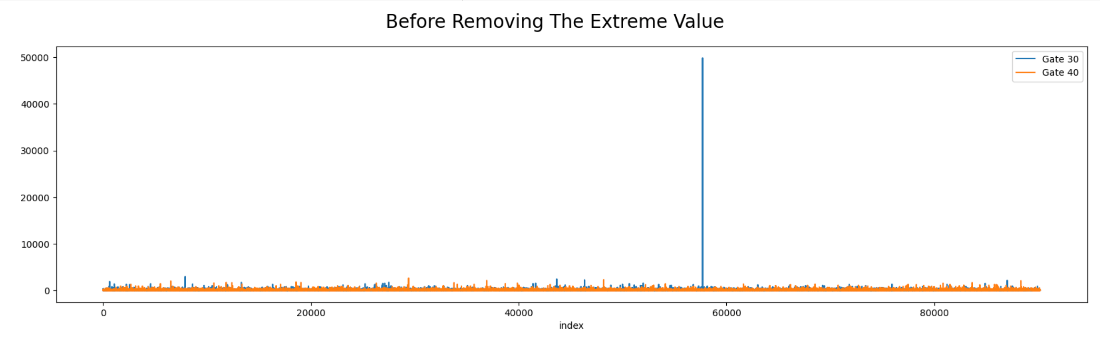


Рис. 3.36. Демонстрация выброса

Визуального анализа часто бывает не достаточно, воспользуемся методом детекции аномалий под названием «Z - оценка»(рис 3.37-3.42).



Рис. 3.37. Подсчет статистик

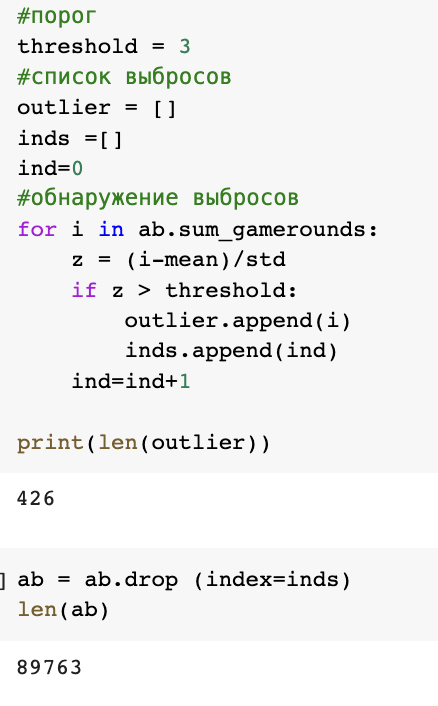


Рис. 3.38. Удаление выбросов из выборки

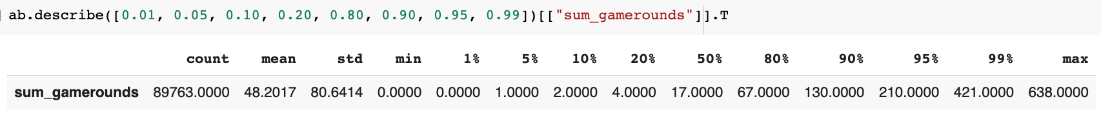


Рис. 3.39. Демонстрация статистик

Как видно из таблицы, максимальное значение сейчас составляет 638 (ранее около 5000).

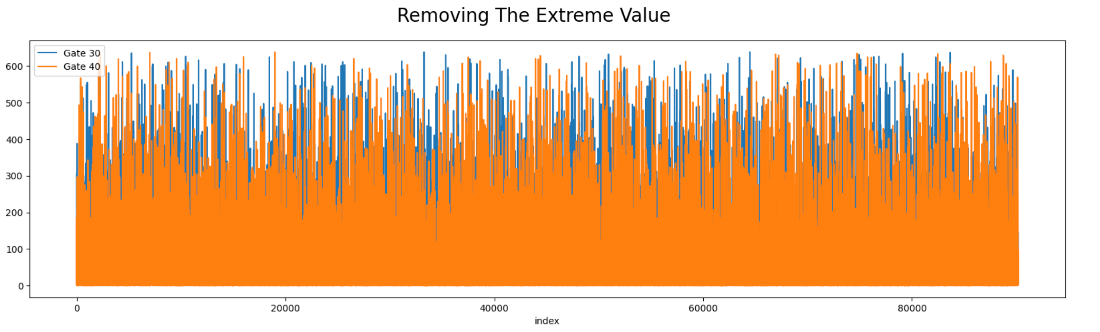


Рис. 3.40. Визуализация данных после детекции аномалий

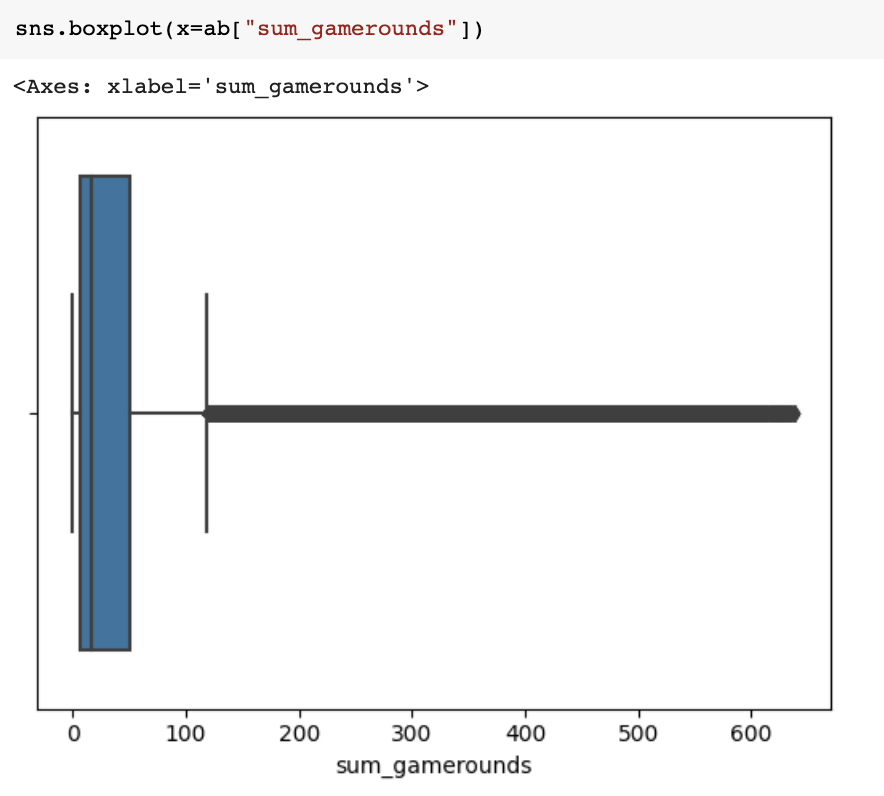


Рис. 3.41. Визуализация данных после детекции аномалий

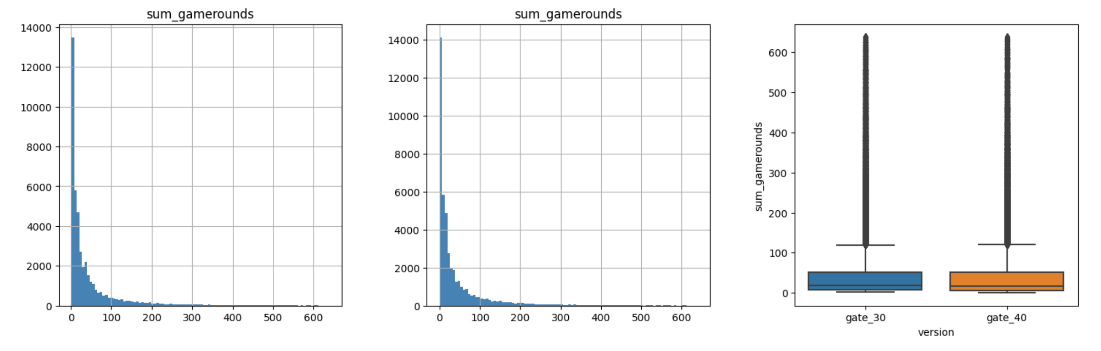


Рис. 3.42. Визуализация данных после детекции аномалий

Пользователи установили игру, но некоторые никогда не играли в нее, они отмечены значением 0 в столбце «sum\_gamerounds». Большинство пользователей играли в игру на ранней стадии и не продвинулись. Проведение исследований и сбор данных об игре и пользователях поможет понять, почему происходит отток пользователей.

Для лучшего понимания визуализируем ответы на вопросы: «Сколько пользователей играло в игру, рассмотрим все раунды?» и «Сколько пользователей находится в раундах с 0 по 200?» (рис 3.43).

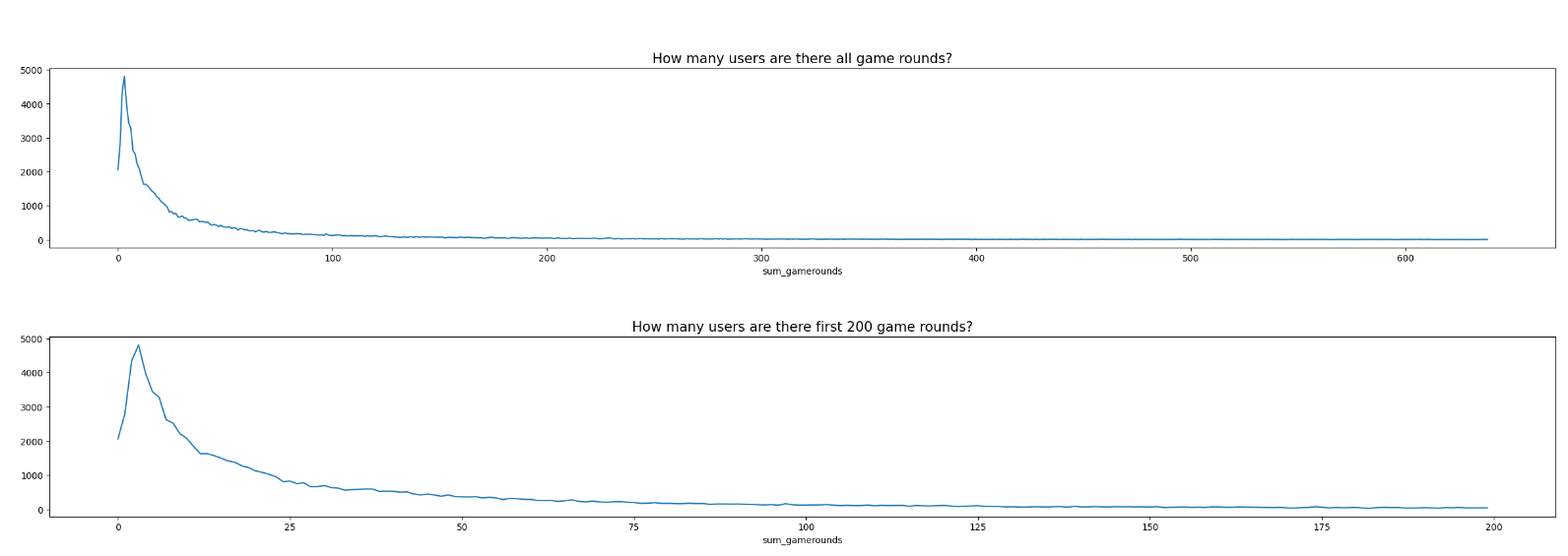


Рис. 3.43. Визуализация данных

Теперь уберём пользователей, не игравших ни один раунд (рис 3.44).

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рис. 3.44. Удаление пользователей, не игравших ни один раунд\

Переменные Retention дают информацию об удержании игроков (рис 3.45).

Retention \_1 - вернулся и играл через 1 день после установки

Retention \_7 - вернулся и играл через 7 дней после установки

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Рис. 3.45. Демонстрация соотношения переменных удержания

Глядя на сводную статистику переменных удержания по версиям и сравнивая с sum\_gamerounds (рис 3.46), можно увидеть сходство между группами. Однако будет полезнее увидеть, есть ли статистически значимая разница.

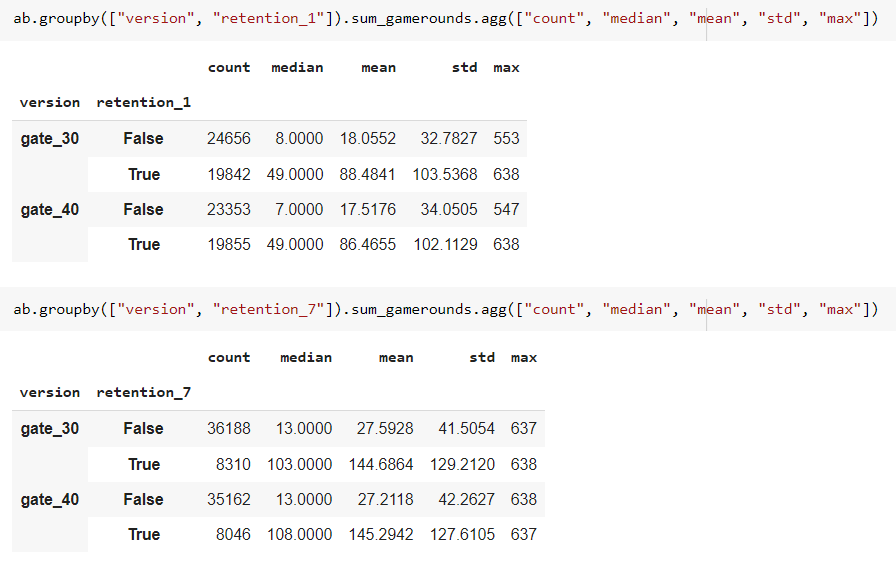


Рис. 3.46. Демонстрация статистик переменных удержания с группировкой

Когда переменные удержания объединяются и две группы сравниваются, сводная статистика здесь также одинакова (рис 3.47).

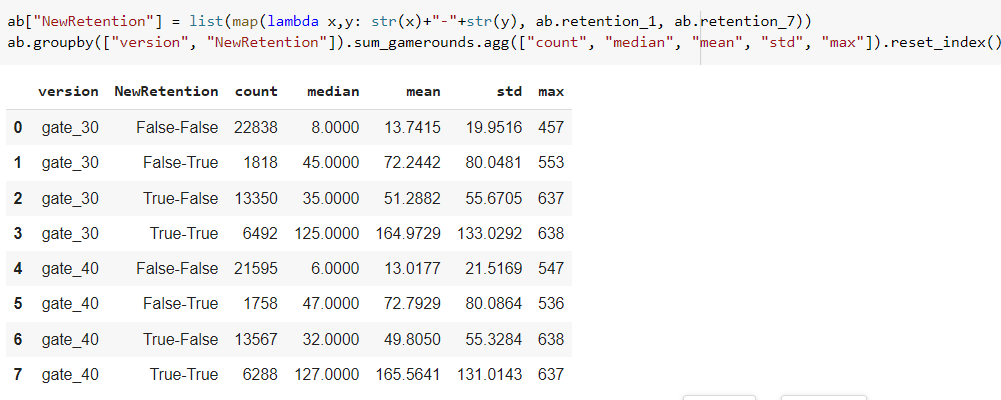


Рис. 3.47. Сводные статистики

Судя по графику, исследованию подвергаются данные, распределение которых отличается от нормального, проведём статистические тесты подходящие под такой размер выборки, для того чтобы в этом убедиться (рис. 3.48).

Изображение выглядит как текст, снимок экрана, Шрифт, документ

Автоматически созданное описание

Рис. 3.48. Тесты на нормальность распределения

Оба теста подтвердили гипотезу, что распределения не являются нормальными, воспользуемся функцией для сравнения выборок

Для определения статистической значимости воспользуемся ранее рассмотренной функцией (рис 3.49).

Изображение выглядит как текст, снимок экрана, Шрифт, линия

Автоматически созданное описание

Рис. 3.49. Результаты теста на различие в выборках

Использовался непараметрический тест, который показал наличие статистически значимой разницы в выборках.

По мере прохождения игры игроки будут сталкиваться с рекламными баннерами, которые заставят их ждать некоторое время, прежде чем они смогут пройти или совершить покупку в приложении или продолжить игру. В этом проекте проанализированы результаты A/B-тестирования, когда первый баннер в Cookie Cats была перемещён с 30-го на 40-й уровень. Особое внимание обращено на удержание игроков и количество игровых раундов.

После применения A/B-тестирования результат анализа даёт важную информацию. Shapiro Testing отклонил H0 из-за предположения о нормальности. Поэтому нужно было применить непараметрический тест Манна-Уитни U для сравнения двух групп. В результате Mann Whitney U Testing отклонил гипотезу H0, группы A/B не похожи.

Вкратце, между двумя группами существует статистически значимая разница в перемещении первого рекламного баннера с уровня 30 на уровень 40. Осталось выяснить какую из версий можно признать “победителем”, для этого проанализируем средние значения удержания игроков (рис 3.50).

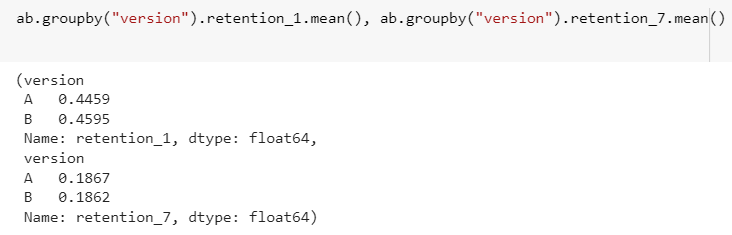


Рис.3.50 Результаты удержания игроков для различных версий.

В качестве вывода можем сформулировать следующее утверждение: для удержания игроков играющих в игру спустя один день после установки важно перемещение рекламных баннеров с 30 на 40 уровень, Для игроков играющих в игру спустя неделю после её установки размещение первого рекламного баннера на 30 или 40 уровне не имеет существенного значения

# Заключение

В результате проведенного исследования были рассмотрены различные методы статистического анализа для A/B тестирования, а также методы прогнозирования будущих результатов. Были проанализированы два подхода - байесовский и частотный - на искусственных и реальных данных.

В результате анализа было выявлено, что оба подхода могут быть эффективно использованы для A/B тестирования. Байесовский подход дает возможность учитывать априорные знания и улучшает точность прогнозирования, особенно при работе с небольшими выборками. Частотный подход является более классическим и широко используется в бизнес-аналитике.

Однако, для достижения достоверных результатов необходимо правильно очищать и предобрабатывать данные перед анализом. Также, важным этапом A/B тестирования является оценка статистической значимости полученных результатов и прогнозирование их будущих значений. В данном исследовании были рассмотрены различные методы для достижения этих целей.

Результаты исследования могут быть полезными для бизнес-аналитиков и маркетологов при принятии правильных решений, улучшении эффективности маркетинговых кампаний и продуктов. Однако, необходимо учитывать, что каждый случай A/B тестирования является уникальным и может требовать индивидуального подхода к анализу и прогнозированию результатов.

В целом, результаты исследования подтверждают актуальность темы A/B тестирования и важность правильного статистического анализа и прогнозирования результатов для достижения успешных бизнес-результатов.

# Список литературы

1. T-критерий для независимых выборок // ibm.com URL: https://www.ibm.com/docs/ru/spss-statistics/26.0.0?topic=tests-independent-samples-t-test (дата обращения: 05.05.2023)
2. Лемешко Б.Ю., Рогожников А.П. Исследование особенностей и мощности некоторых критериев нормальности. - № 4 изд. - Метрология, 2009. - 40 с.
3. На что мы обращаем внимание при расчете статистической значимости A/B-теста // habr.com URL: https://habr.com/ru/companies/uchi\_ru/articles/500918/ (дата обращения: 10.05.2023)
4. aByes paket // github.com URL: https://github.com/cbellei/abyes (дата обращения: 10.05.2023)
5. Bayesian A/B Testing with Python: the easy guide // towardsdatascience.com URL: https://towardsdatascience.com/bayesian-a-b-testing-with-python-the-easy-guide-d638f89e0b8a (дата обращения: 10.05.2023).
6. Bayesian A/B Testing: a step-by-step guide // claudiobellei.com URL: http://www.claudiobellei.com/2017/11/02/bayesian-AB-testing/# (дата обращения: 05.05.2023)