Статистика в анализе данных

A/B тестирование
Проверка гипотез с помощью статистических методов

Папахристу Александра, Чан До Минь Чау, Шварцер Мария, Швецова Николь, Ли Александра

Раздел №1:

EDA и статистические тесты

EDA и статистические тесты

Бизнес-задача: провести анализ рынка аренды недвижимости США

- **Наш заказчик - предприниматель, собирающийся выйти на американский рынок со своим риелторским агентством.**
- **Предметная область - рынок аренды недвижимости в США.**
- **Датасет содержит информацию об объявлениях об аренде за 2018-2019 года с ключевыми характеристиками объектов**

Немного о датасете:

- **100** тысяч данных о предложениях **2018-2019** годов
- Самый популярный тип аренды - **apartment** (**99%** от всех объявлений)
- Все цены указаны в долларах
- Периодичность оплаты - каждый месяц (в основном)
- В основном данные с платформы **RentDigs.com**

В объявлениях указана цена, площадь, краткое описание, удобства, местоположение и отношение к животным арендодателя и другие признаки

Источник: https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified

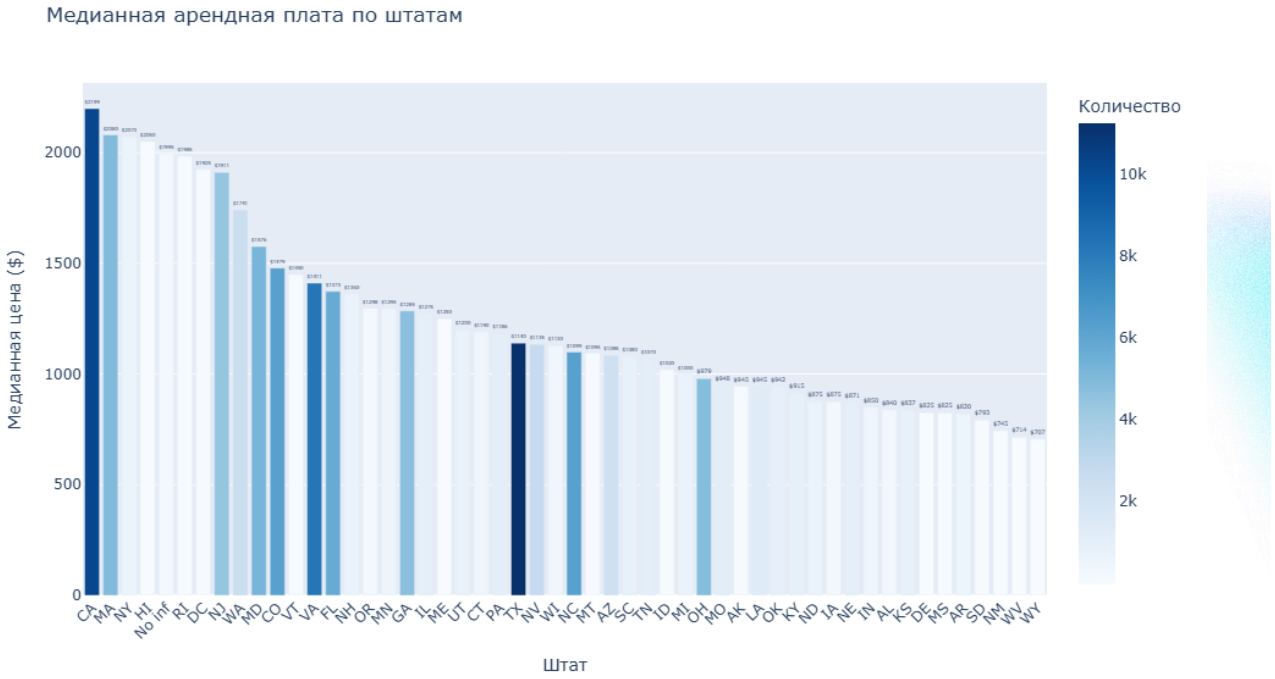
84% объявлений содержат детализированную информацию об удобствах, что свидетельствует о высо кой значимости данного параметра для конкуренции на рынке аренды

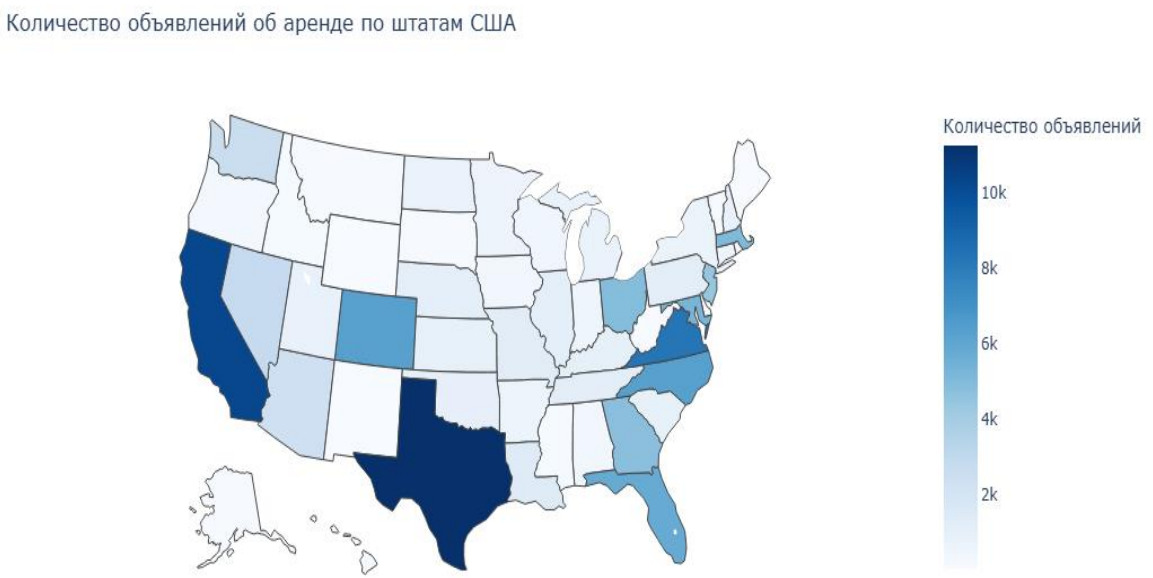
ТОП-5 удобств (по частоте упоминания):

- **Парковка (43,729) — абсолютный must-have**
- **Бассейн (43,425) — почти равен парковке по важности**
- **Тренажерный зал (37,227) — стандарт для многоквартирных домов**
- **Патио/Терраса (26,460) — важное outdoor - пространство**
- **Стиральная машина + сушилка (25,922) — бытовая необходимость**

Эти 5 удобств формируют базовый стандарт рынка

Ценовой аспект наших данных

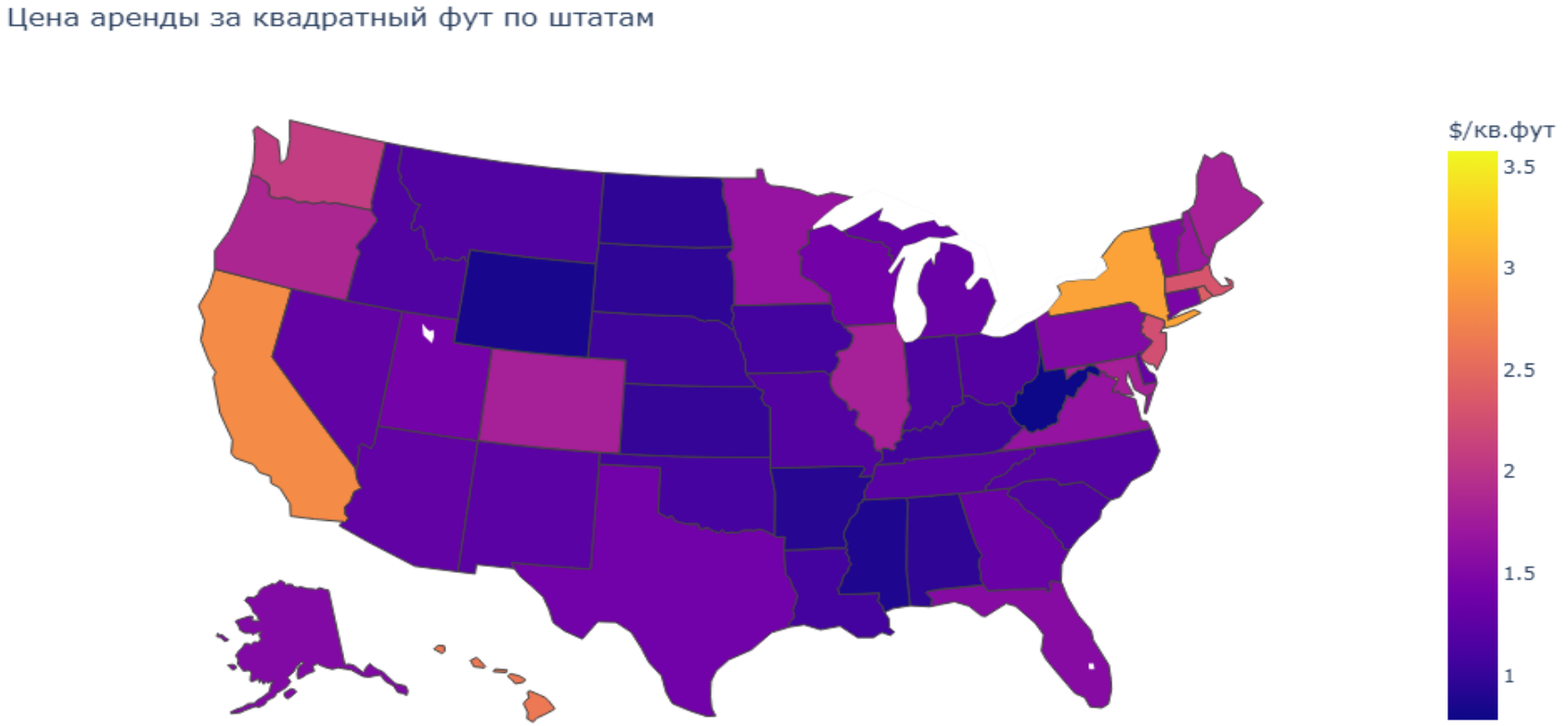


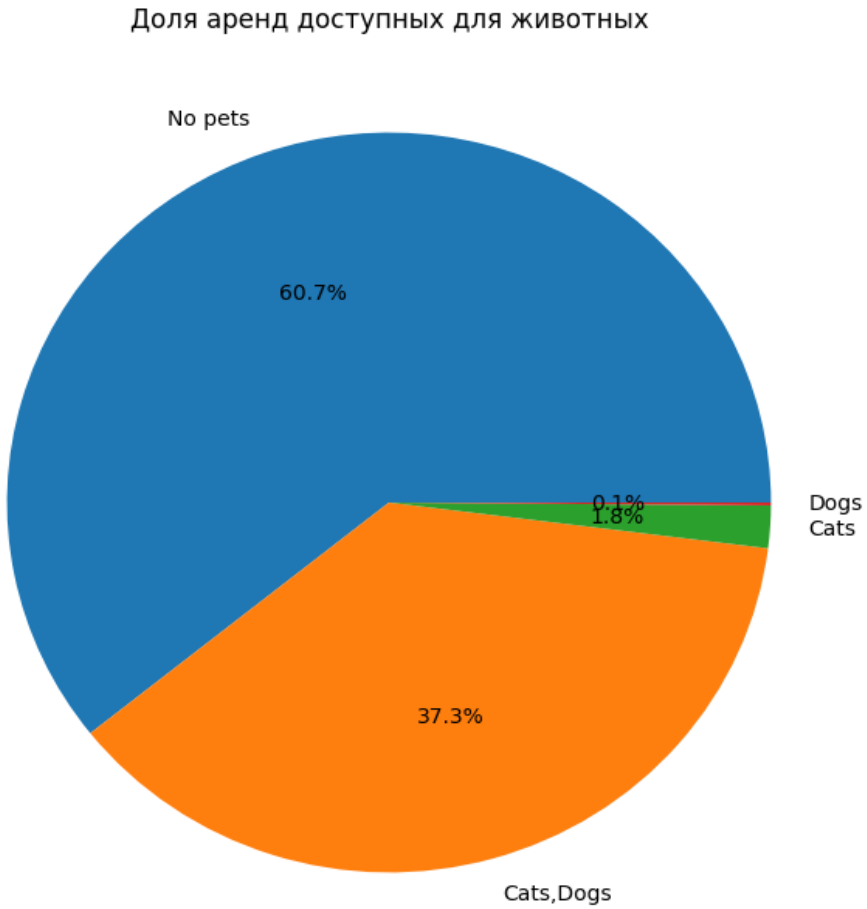


Самые популярные штаты по количеству объявлений: **Ка лифорния** (10к объявлений), **Техас** (11к объявлений), **Ви ргиния** (8к объявлений)
Они являются самыми активными конкурентными рынк ами.

Также **Калифорния, Массачусетс** и **Нью-Йорк** - топ 3 штата по медиан ной арендной плате. Именно там с амое дорогое жилье

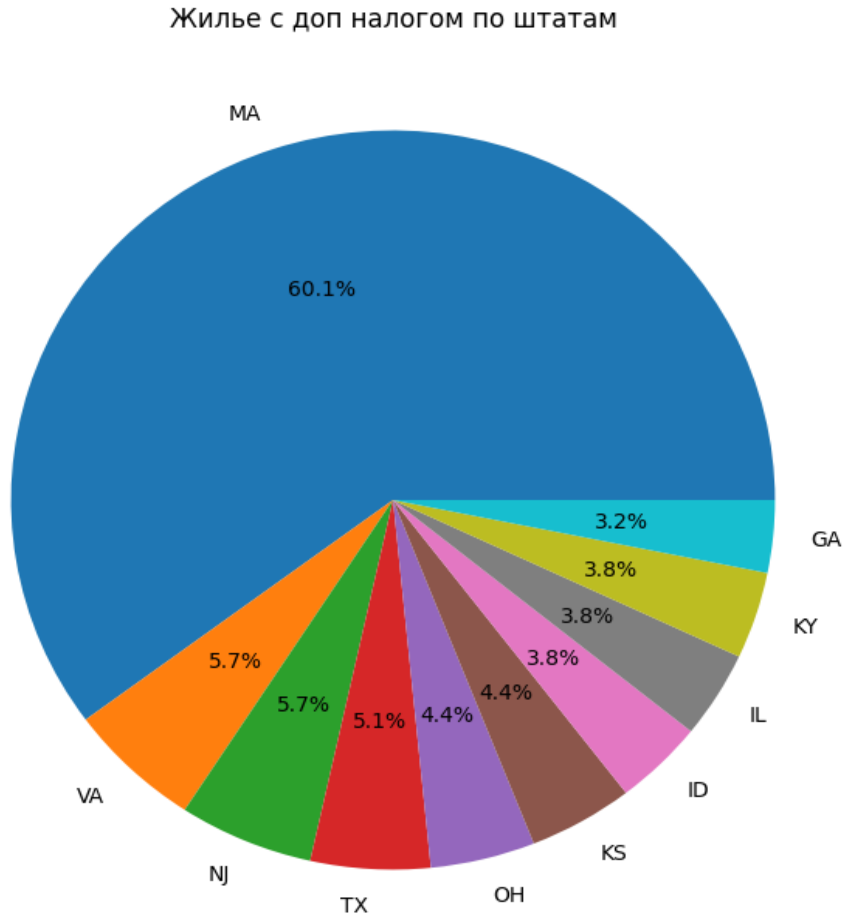
Создадим новый признак - нашу с целевую переменную — цена за квадратный фут.
Это ключевой этап подготовки данных, так как цена за квадратный метр является нормализова нным показателем, который позволяет объективно сравнивать объекты недвижимости разной площади и стоимости.





Большая доля запрета на животных может со здавать барьер для потенциальных арендода торов

Большинство квартир с налогом сверху в Массачусетсе, а именно в Бостоне. Это может быть связано с расположением университета **M IT**



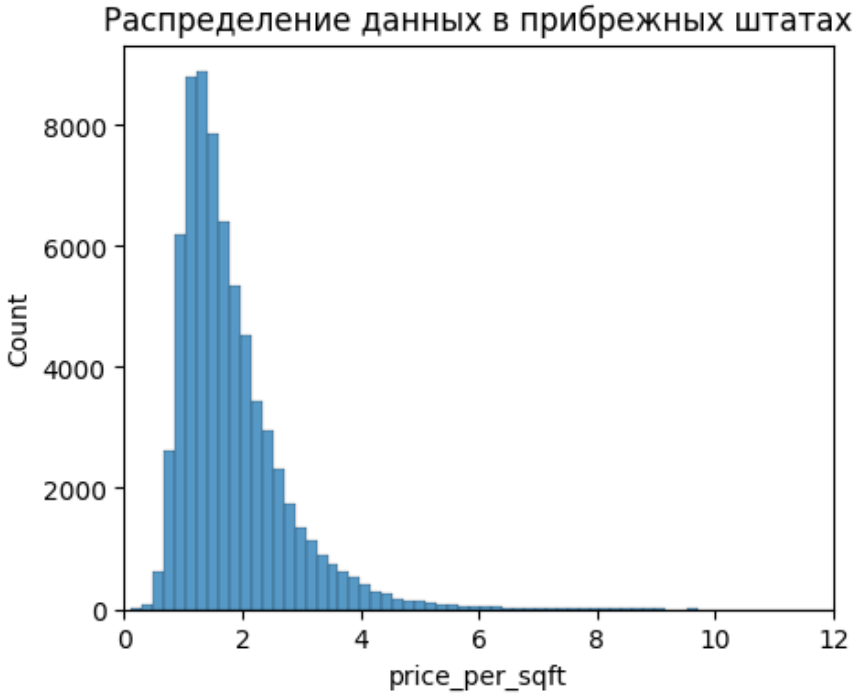
Прибрежные штаты и внутренние

Разделим всю территорию Америки на 2 группы и сравним их

Гипотеза на основе EDA:

H0: Распределения цен аренды за квадратный фут в прибрежных и внутренних штатах идентичны

H1: Распределения цен аренды за квадратный фут в прибрежных и внутренних штатах различаются. (предполагаем, что цены в прибрежных штатах выше.)



Выявили на этапе анализа распределения данных:

Рынок прибрежных штатов:

- **более "элитный" — больше дорогих объектов (разница среднее - медиана больше)**
- **Больше разнообразия — шире диапазон цен (большее std)**
- **Более активный рынок — больше объявлений (возможно, выше оборот)**

Подбор статистического теста и результат

Наши данные:

- Огромные выборки
- Несбалансированные по размеру
- Не нормально распределённые
- Дисперсии не гомогенны
- Независимы
- Есть выбросы

- **Используем - U-критерий Манна-Уитни (пользуемся методом уменьшения выборки для балансировки размера) - получаем статистически значимый результат и отклоняем нулевую гипотезу.**

- **Для оценки размера эффекта результата используем ранг бисериальную корреляцию = -0.796 (первая группа (coast) имеет значительно более высокие значения систематически)**

- **Также нашли Эфронов доверительный интервал для разницы медиан : 0.334 - 0.352**

Бизнес применение:

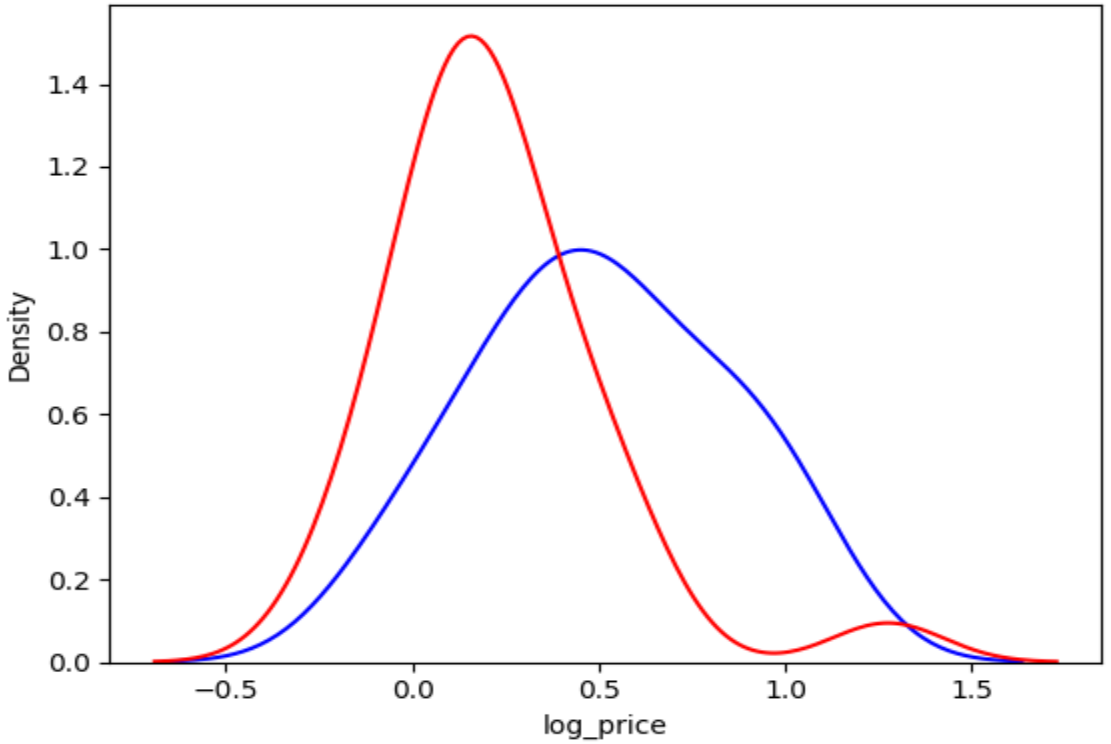
- **Разница в 0.34$ за кв. фут- это не просто статистика, это дополнительный доход, который в среднем мы будем получать дополнительно за кв. фут, продвигая объекты в прибрежных зонах, вместо внутренних**

- **Во внутренних штатах можем делать акцент на ценовую эффективность (для клиентов)**

Теперь переформулируем немного нашу гипотезу:

H0 - Средняя цена за квадратный фут недвижимости в прибрежных штатах и средняя цена во внутренних штатах не отличаются
H1 - Средняя цена за квадратный фут недвижимости в прибрежных штатах превышает среднюю цену во внутренних штатах.

Рассмотрим распределение логарифмированных средних показателей по штатам в каждой выборке



Наши данные:

- Одна выборка распределена нормально, вторая почти нормально
- Дисперсии гомогенны
- Выборки несвязные

- Применяем **t**-тест Стьюдента и получаем статистически значимый результат - отклоняем **H0**

Общий вывод:
Прибрежные штаты однозначно дороже внутренних, причём это проявляется ка к на уровне средних значений, так и на уровне общего распределения цен.

- **Не просто "в среднем" дороже (t-test), а систематически и стабильно дорож е (Mann-Whitney)**
- **Разница не случайна — подтвержден а двумя независимыми методами**
- **Эффект устойчив — проявляется при разных способах измерения**

Значимость для бизнеса:

- **Создать 2 отдельных отдела/экспертов:**
  - **Внутренние штаты: эксперты по семейному ж илью**
  - **Прибрежные штаты: эксперты по премиум-се гменту**
- **Разные скрипты продаж для разных регионов**
- **Корпоративная аренда:**
  - Размещать сотрудников: во внутренних шт атах выгоднее на 35%
  - Офисные помещения: внутренние штаты д ают экономию на аренде
- **Вывод:**
- **Для пассивного дохода — внутренние штаты.**
- **Для роста капитала + доход — прибрежные.**

Техас **VS** Калифорния

Сравнительный анализ рынка недвижимости в Калифорнии и Техасе

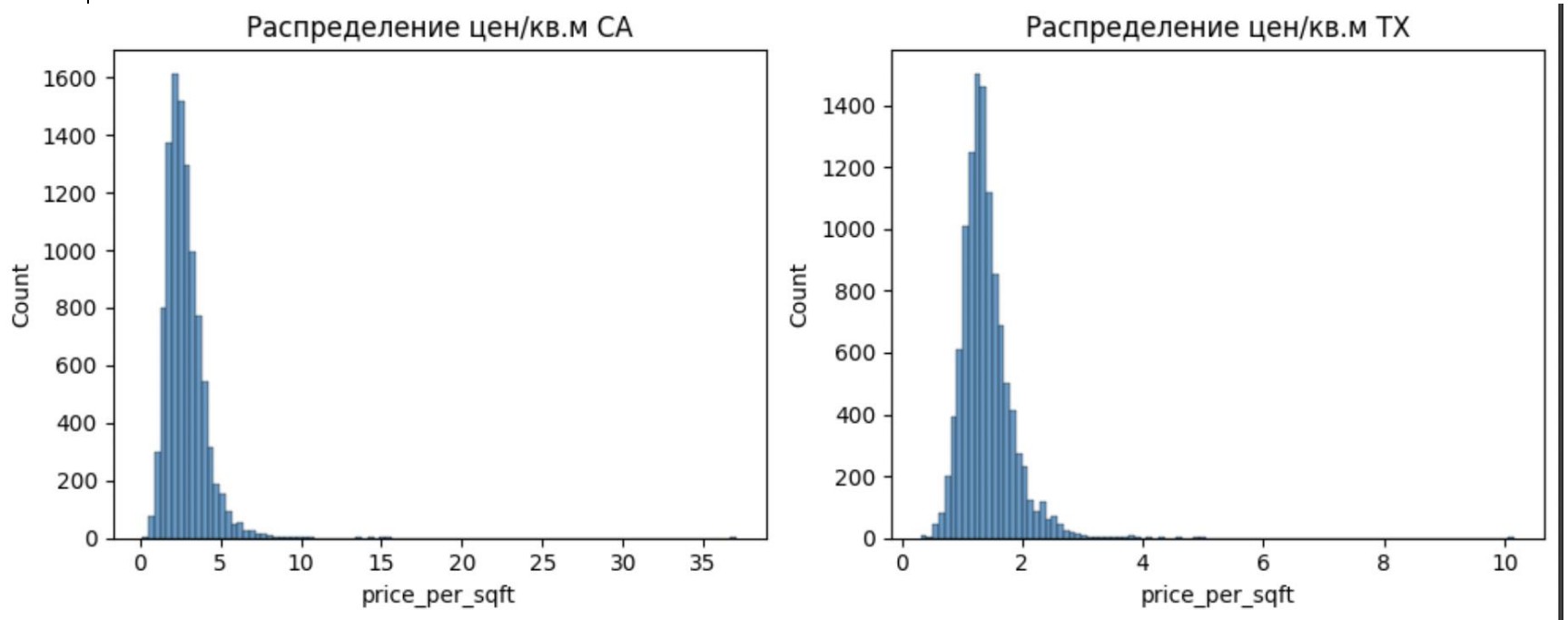
Техас VS Калифорния

Как мы выяснили ранее: Калифорния и Техас я вляются одними из самых активных рынков а ренды, количество объявлений превышает 10к

Цель: выявить, какой штат является более "до ступным" для риелторской компании

Оба распределения имеют правостороннюю асс иметрию. Основная масса наблюдений у CA в ди апазоне 1-4 и более длинный хвост, у Техаса 1-2.

Также выяснилось, что дисперсия цены за фут в Калифорнии в 9 раз больше чем в Техасе



Гипотеза: Рынок недвижимости Техаса более стандартизирован по цене за фут, чем рынок Калифорнии

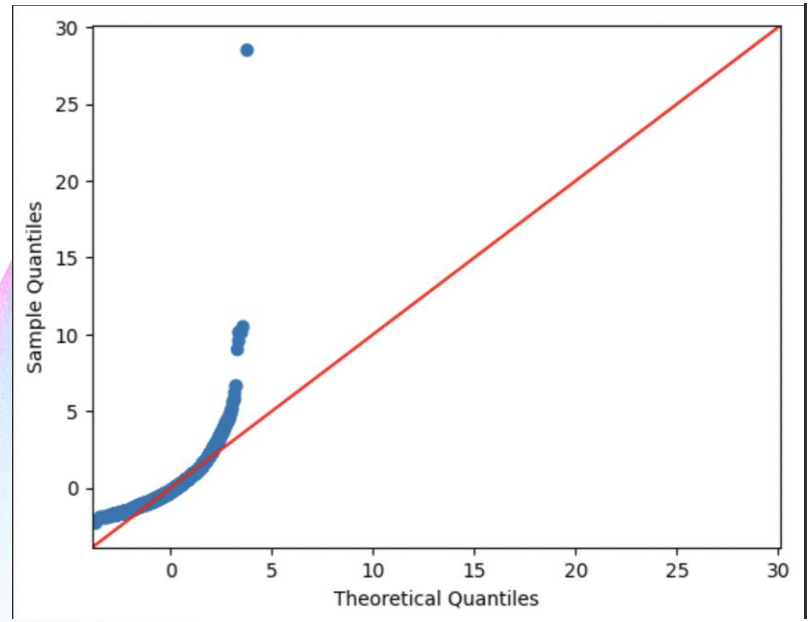
Результаты **EDA** показали, что:

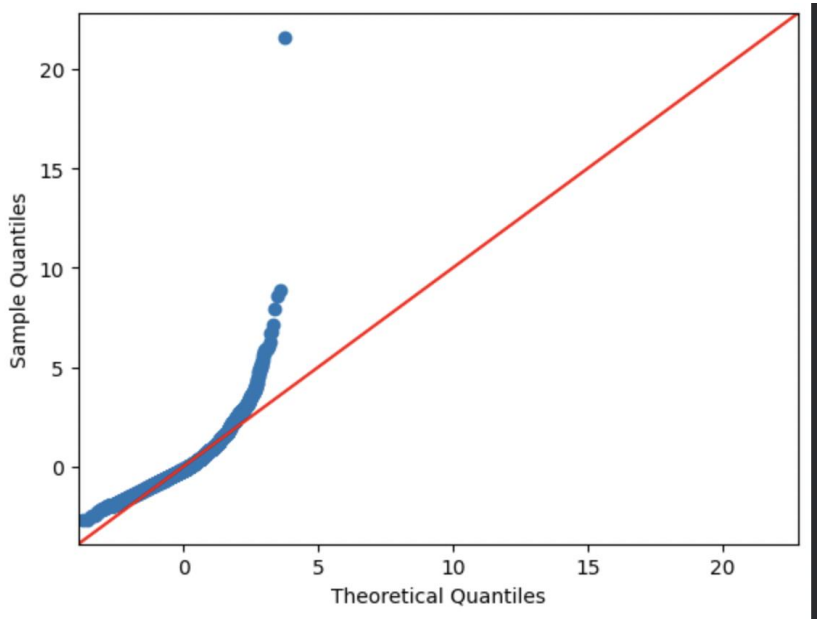
- Для Калифорнии наблюдается более широкий диапазон значений и длинный правый хвост -> Большое количество экстремально дорогих объектов
- Рынок Техаса имеет более компактное распределение цен, разброс меньше, выпады менее выражены (однако они все равно есть)

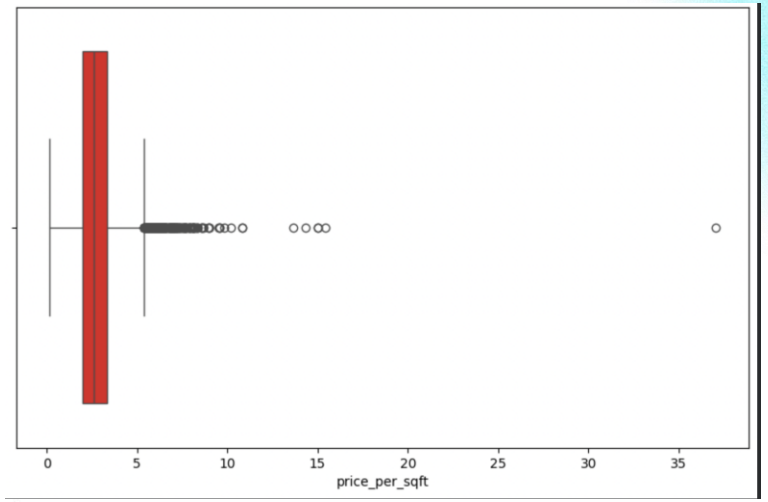
Таким образом, исходя из выявленных различий есть предпосылки выдвинуть гипотезу, что рынок Техаса более стандартизирован по ценовой структуре.

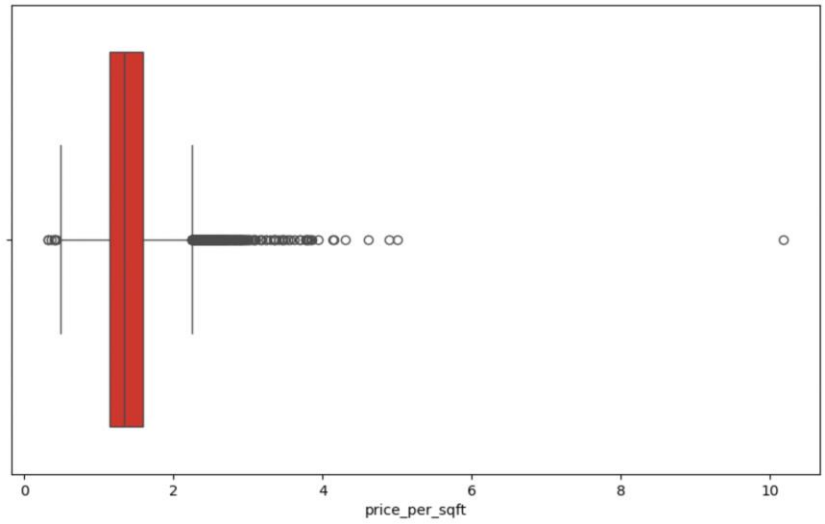
Гипотеза: Рынок недвижимости Техаса более стандар тизирован по цене за фут, чем рынок Калифорнии

Для подтверждения нашей гипотезы воспользуемся ста тистическими тестами. Но для начала необходимо корр ектно выбрать тест









Статистические тесты

Степень стандартизации ценового значения измеряется разбросом вокруг средней цены (то есть дисперсией). Для проверки гипотезы мы будем сравнивать дисперсии цен за фут в двух выборках.

Сформулируем:

H0: Дисперсии цен за фут у Техаса и Калифорнии не имеют статистически значимых различий

H1: Дисперсии цен за фут Техаса и Калифорнии имеют статистически значимое различие

Тест Брауна-Форсайта

```
Brown-Forsythe statistic = 3993.002659077937
p-value = 0.0
```

Тест Флингера-Килиана

```
Статистика Флигнера-Килиана: 5260.915012913338
p-value: 0.0
```

Оба теста показали **p-value <0,05,** поэтому гипотеза **H0** отклоняется

Также соотношение коэффицентов вариации у цены в Калифорнии и цены в Техасе равняется **1,48,** что также подтверждает гипотезу о статистическ и значимом различии дисперсий

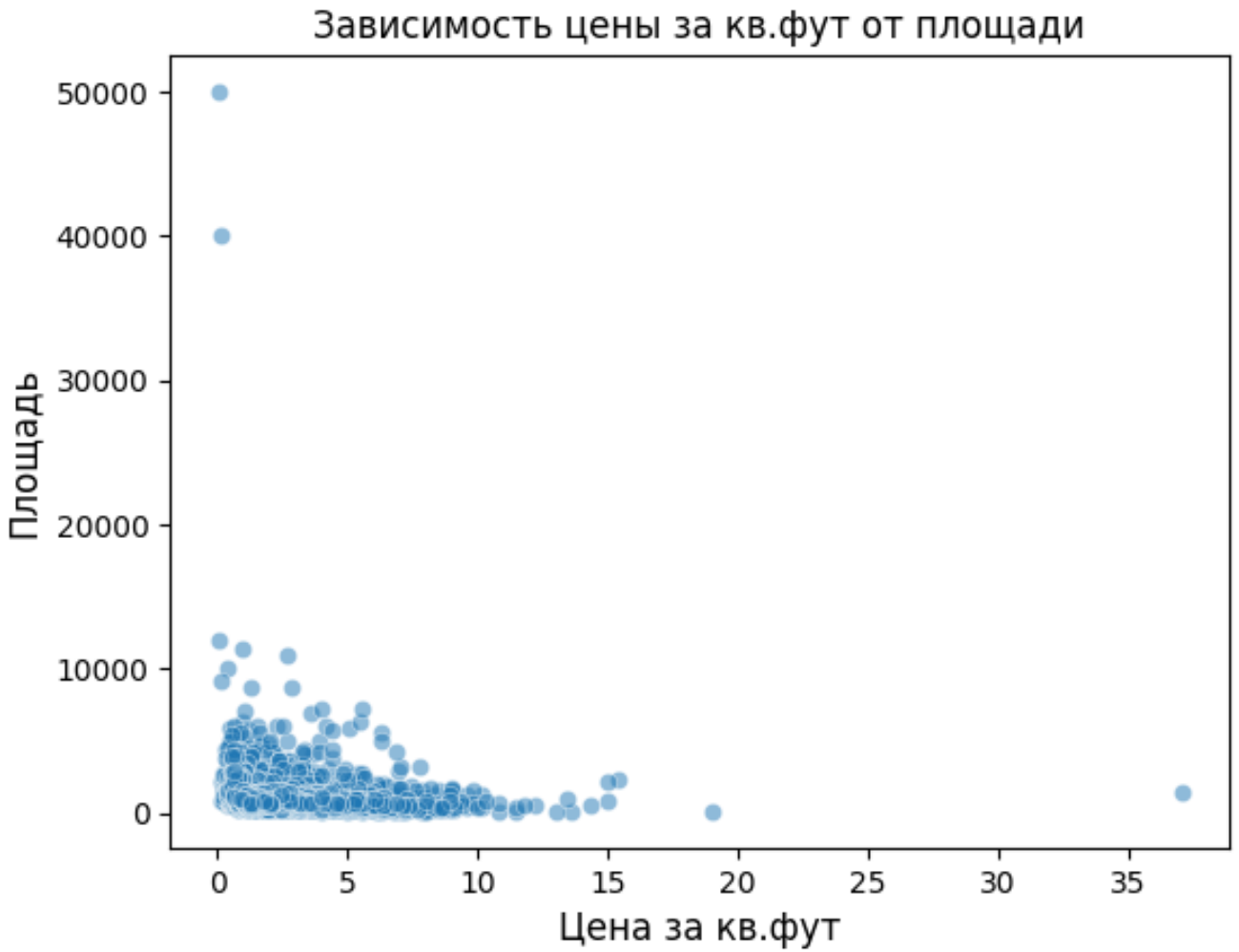
Выводы

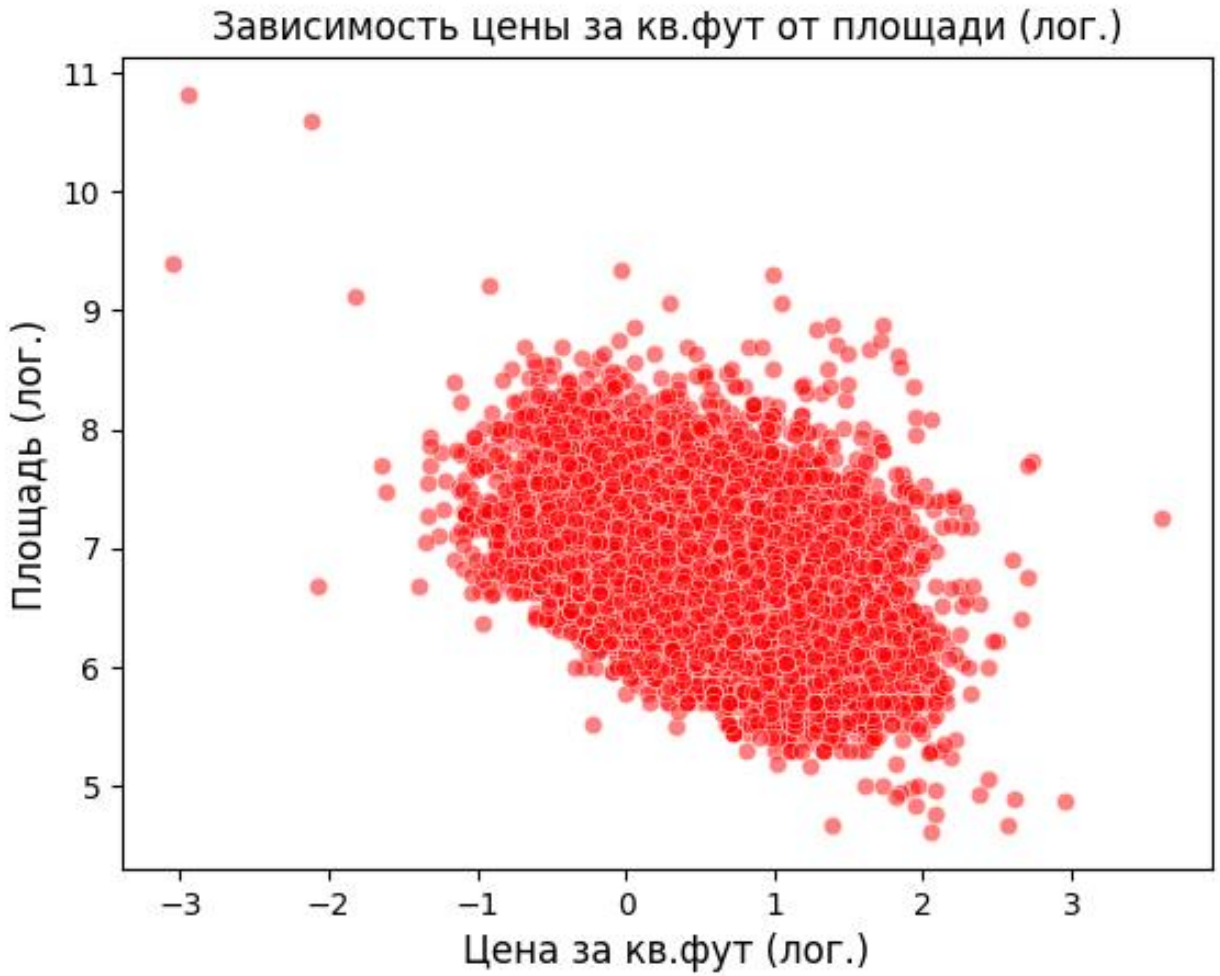
- Дисперсии цены за фут Калифорнии и Техаса значительно различаются
- Рынок аренды Техаса является более стандартизированным. Цена в Калифорнии ведет себя гораздо более непредсказуемо, зависит от множества других факторов,
- Для начинающей риелторской компании лучше обратить внимание на Техас, поскольку он более стабилен

Общая площадь и цена за кв. фут

Есть ли связь между ними?

Общая картина на рынке США

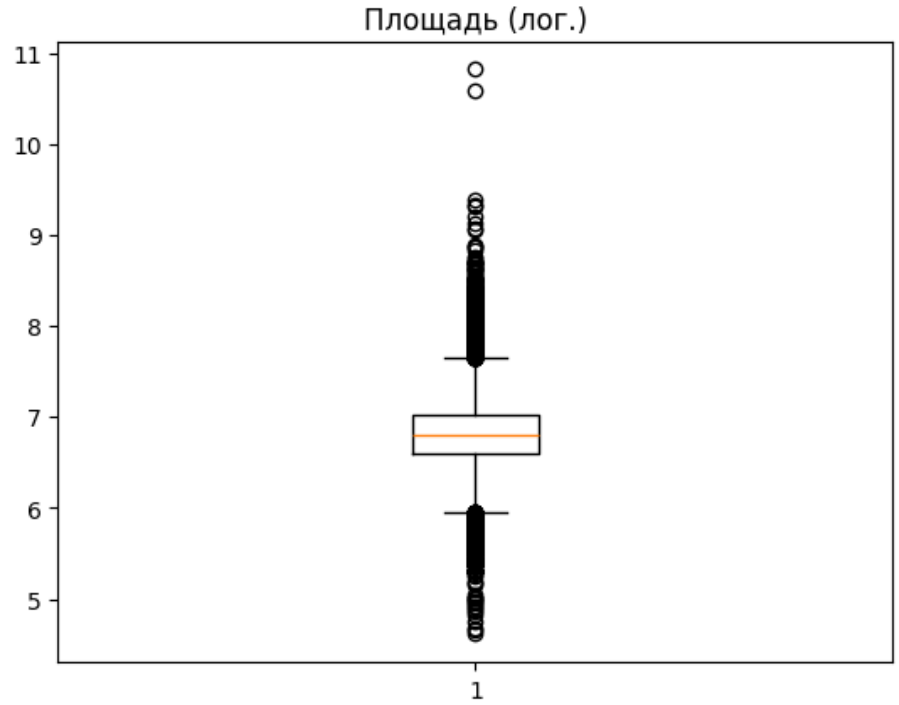


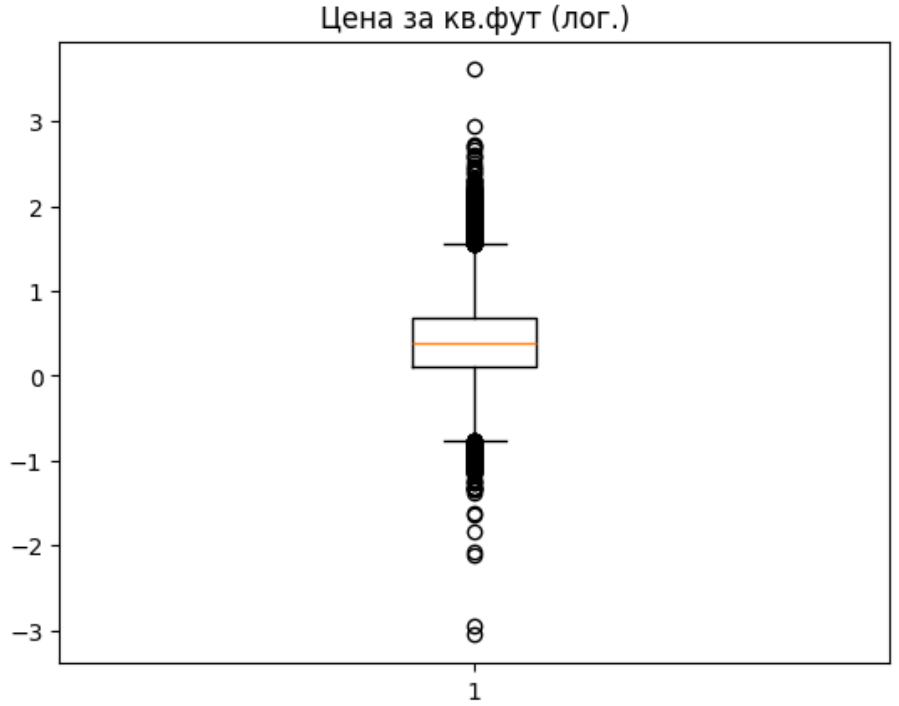


Гипотеза: с увеличением площади цена за кв. фут падает

Тест ранговой корреляции Спирмена:

- Ранги
- Не требует нормального распределения данных
- Устойчив к выбросам
- Непараметрический





Тест ранговой корреляции Кендалла-тау:

- Ранги
- Не требует нормального распределения данных
- Устойчив к выбросам
- Непараметрический

Статистические тесты

Связь между признаками в нашем случае - наличие з начимой монотонной зависимости

Наши гипотезы**:**
H0: Уменьшение цены за фут не связано с увеличени ем площади жилья
H1: Уменьшения цены за фут связано с увеличением площади жилья

Тест ранговой корреляции Спирмена

```
Spearman's correlation coefficient: -0.34023935331789473
p-value: 0.0
```

Тест ранговой корреляции Кендалла-тау

```
Kendall Rank correlation: -0.23345
p-value: 0.0
```

Оба теста показали **p-value <0,05,** поэтому гипоте за **H0** отклоняется

Выводы

- Подтвердилась значимая умеренная монотонная зави симость между признаками
- Нужно учитывать влияние площади на цену за квадра тный фут при ценообразовании и переговорах с клиен том
- Площадь не является единственным фактором влиян ия, что говорит о необходимости сегментации жилья, чтобы была возможность делать акцент на других пар аметрах жилья

Раздел №2:
A/B тестирование

Проанализируем результаты проведения некоторого **A/B** тестирования в компании из сферы образования

Анализ данных

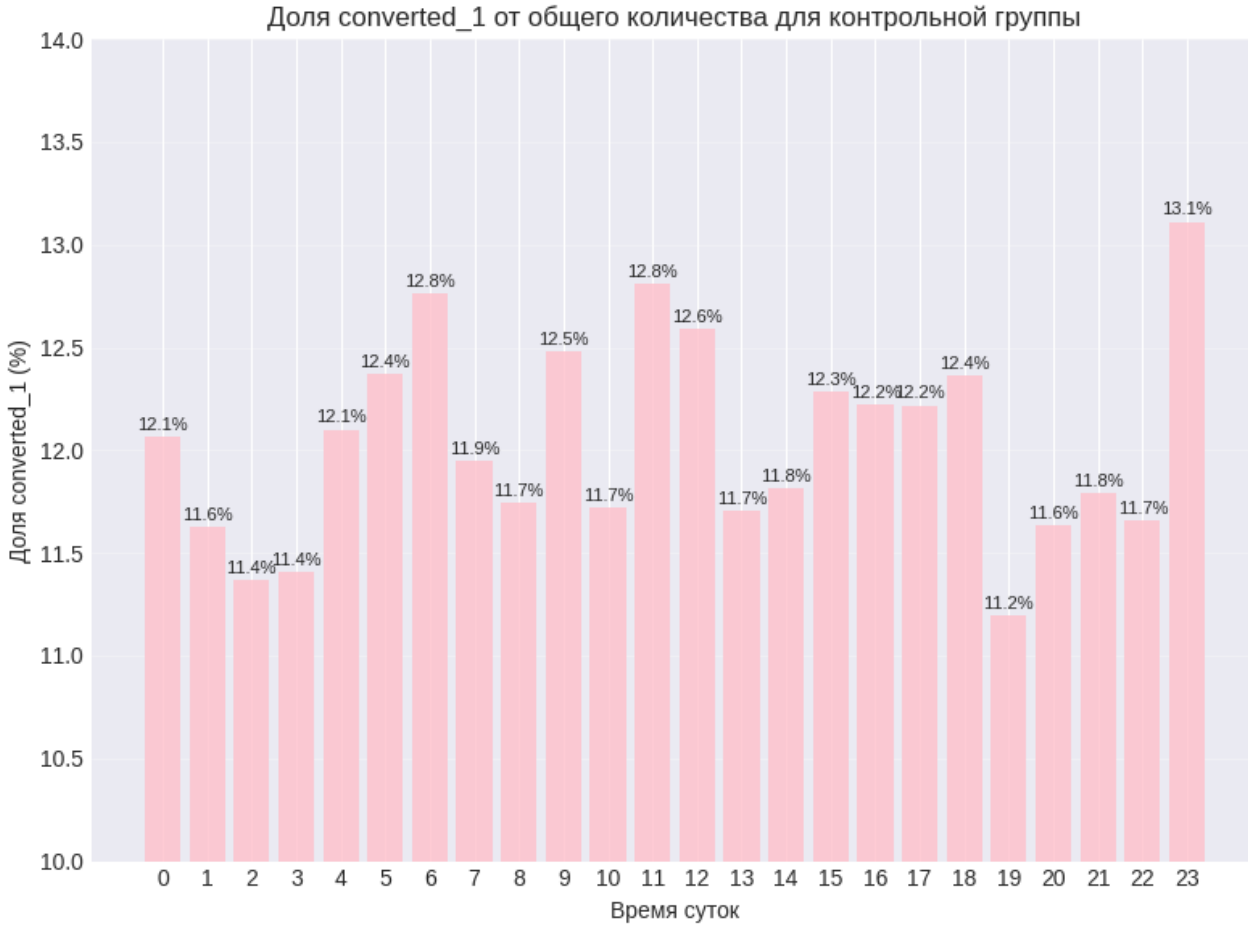
	Контрольная группа	Экспериментальная группа	Всего
Количество данных	**144319**	**144316**	**288635**
Время начала	2025-01-02 13:42:05	2025-01-02 13:42:15	2025-01-02 13:42:05
Время конца	2025-01-24 13:41:54	2025-01-24 13:41:44	2025-01-24 13:41:54

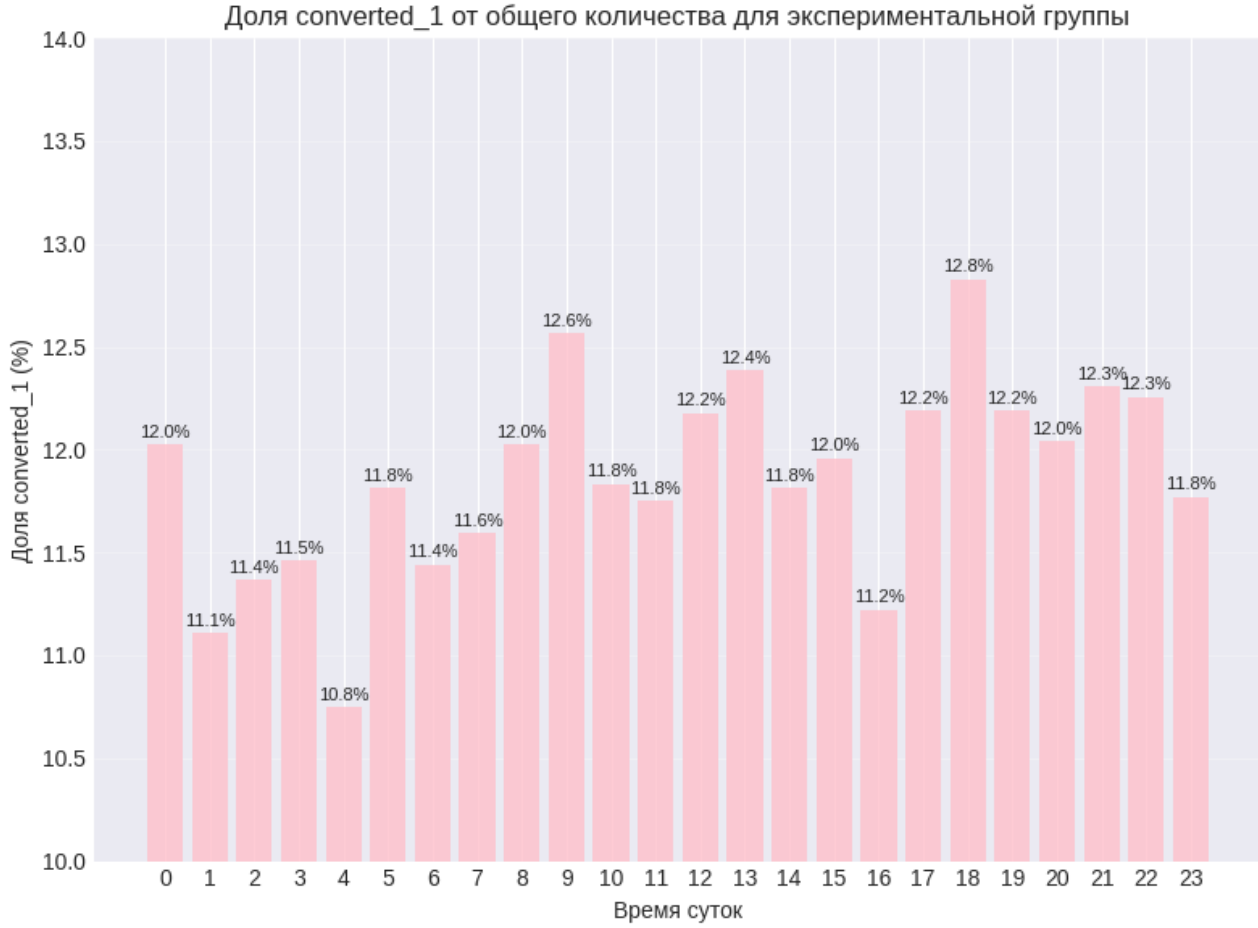
Эксперименты в обоих группах проводились в течение **4** календарных недель

При этом оказалось, что некоторым людям из контрольной группы показывали новую версию, а некотор ым из экспериментальной - старую версию - это проблема! Данну проблему мы решили обработкой каждо го типа дубликатов

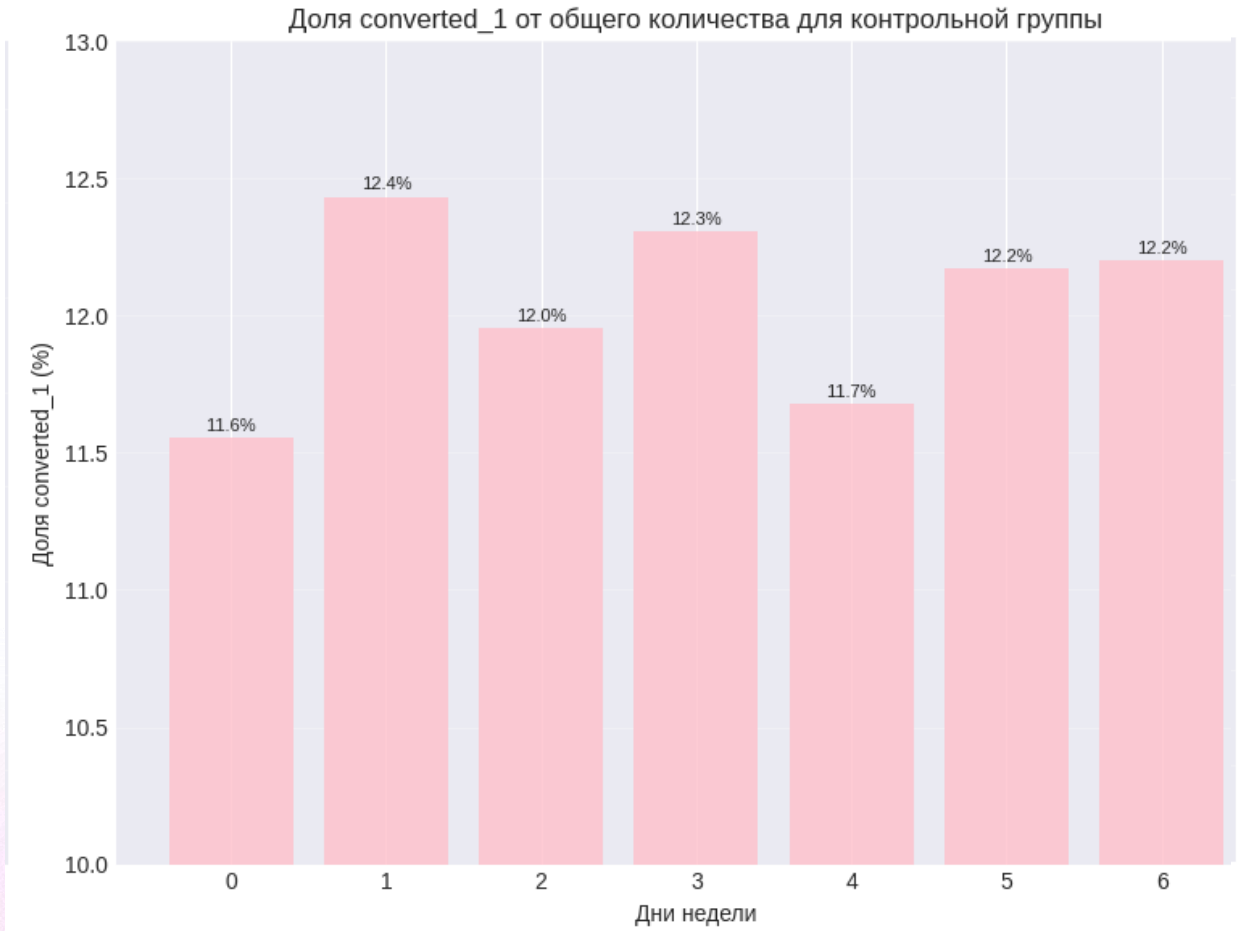
В качестве новых данных мы ввели - время суток эксперимента, день недели, день с начала эксперимента

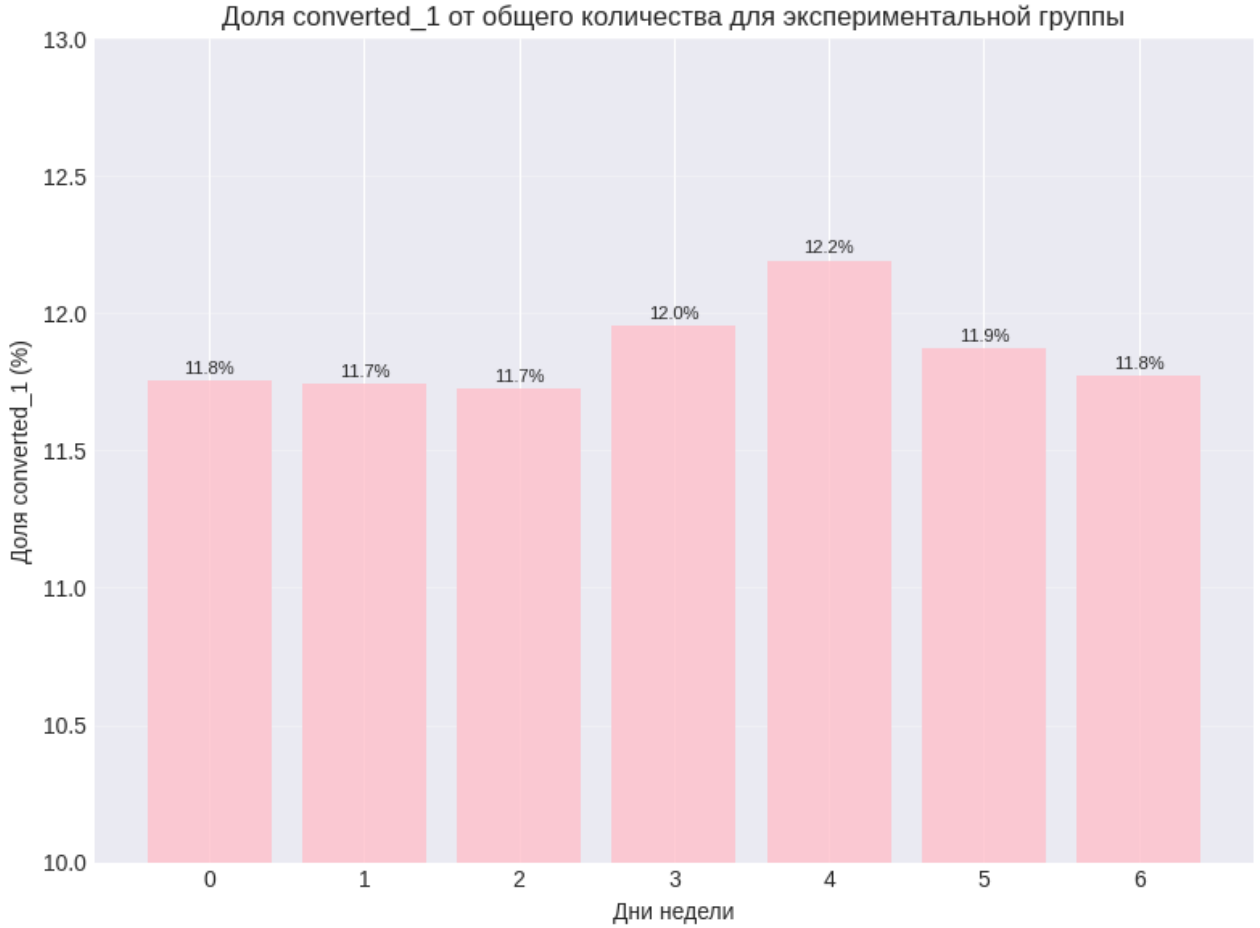
Анализ конверсии по часам



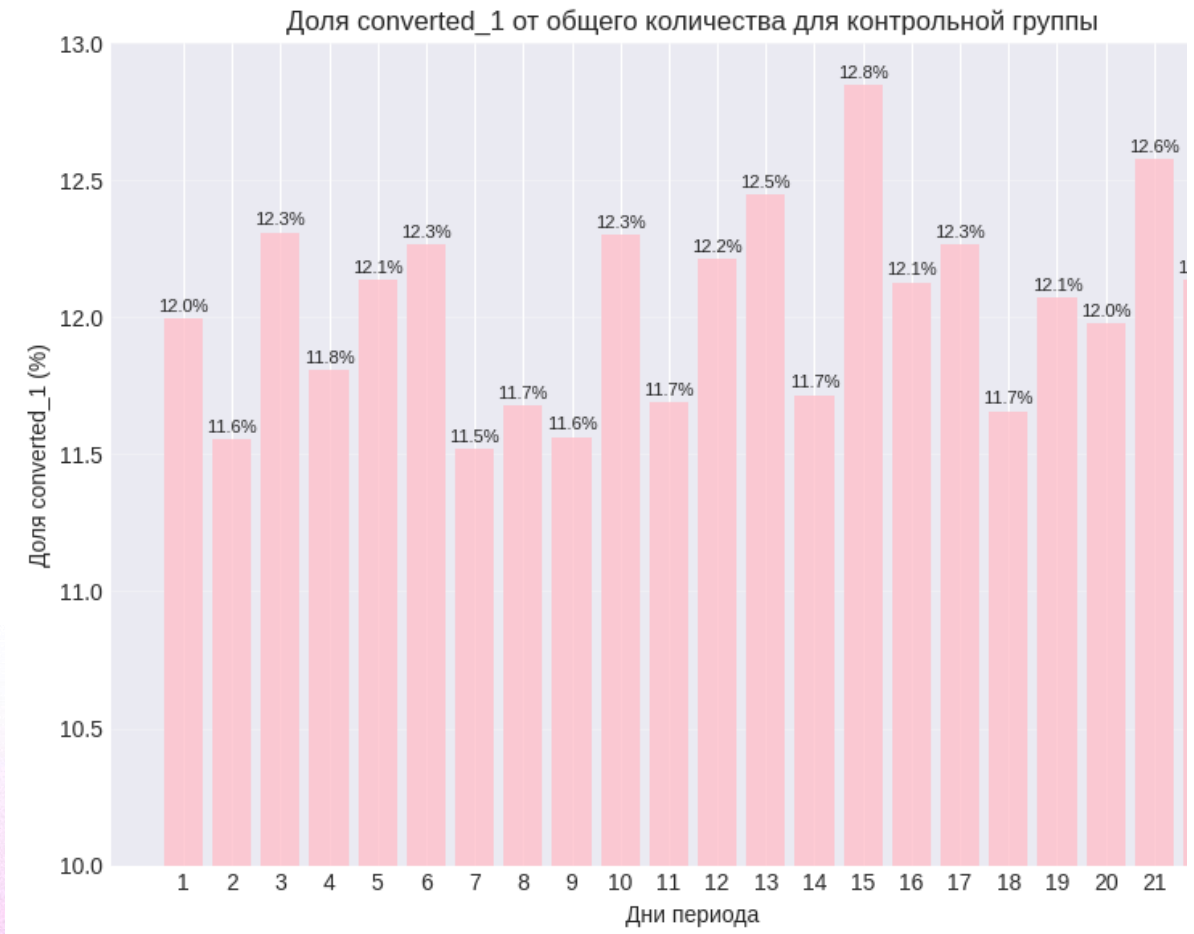


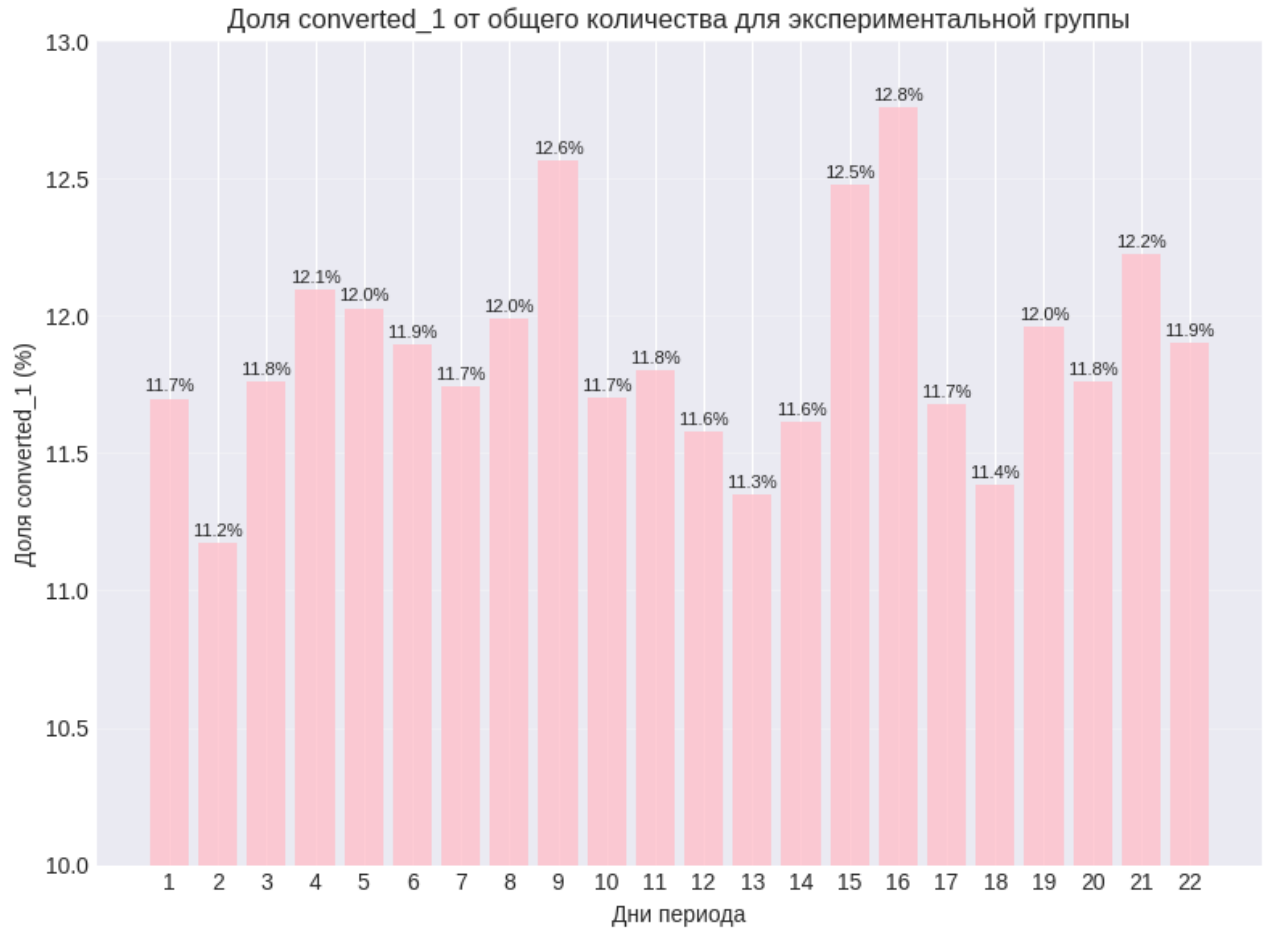
Анализ конверсии по дням недели





Анализ конверсии по дням





Наблюдения

Почасовой анализ:

- **Контрольная группа:** Максимум конверсии (13,1%) в 23:00, минимум (11,2%) в 19:00. Пики активности в 6:00, 11:00 и 23:00

- **Экспериментальная группа:** Максимум (12,8%) в 18:00, минимум в 4:00. Пики активности смещены
- (9:00, 13:00, 18:00)

Анализ по дням недели:

- **Контрольная группа:** Максимум относительной конверсии во вторник (12,4%), минимум в
- понедельник (11,6%)

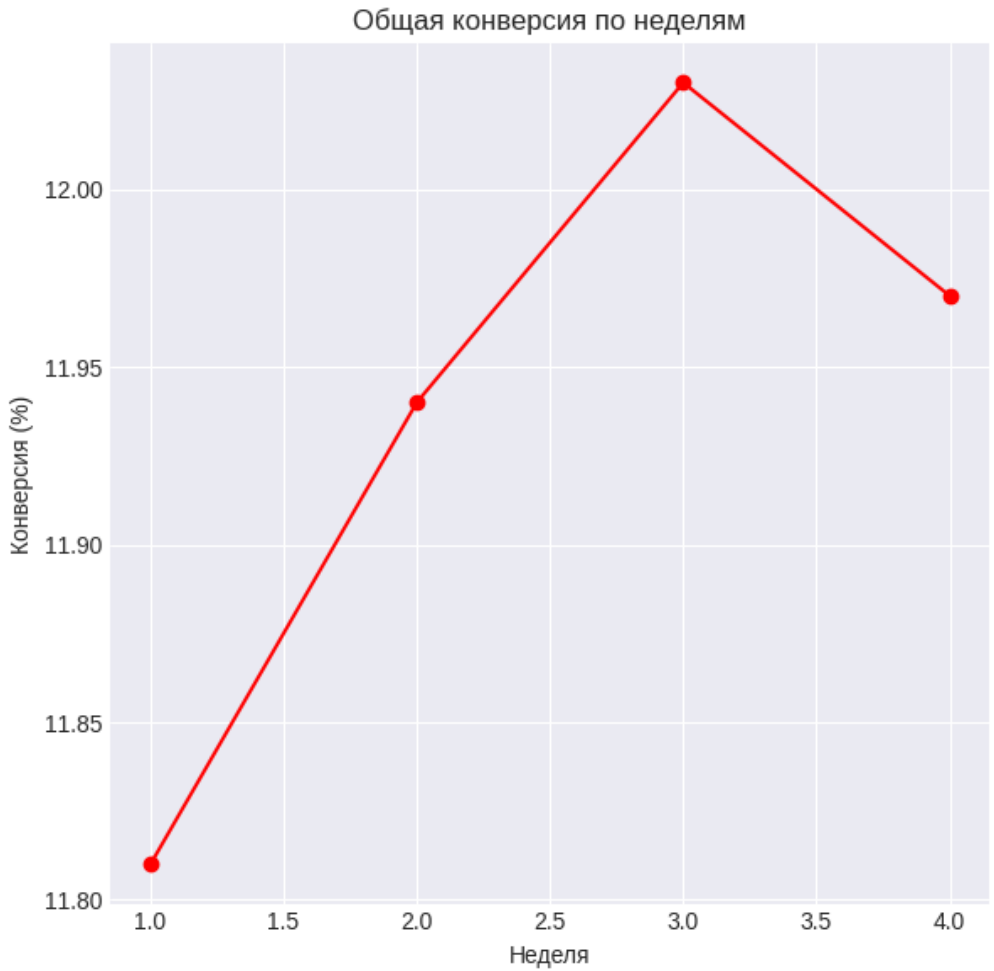
- **Экспериментальная группа:** Максимум в пятницу (12,2%), в остальные дни конверсия стабильна

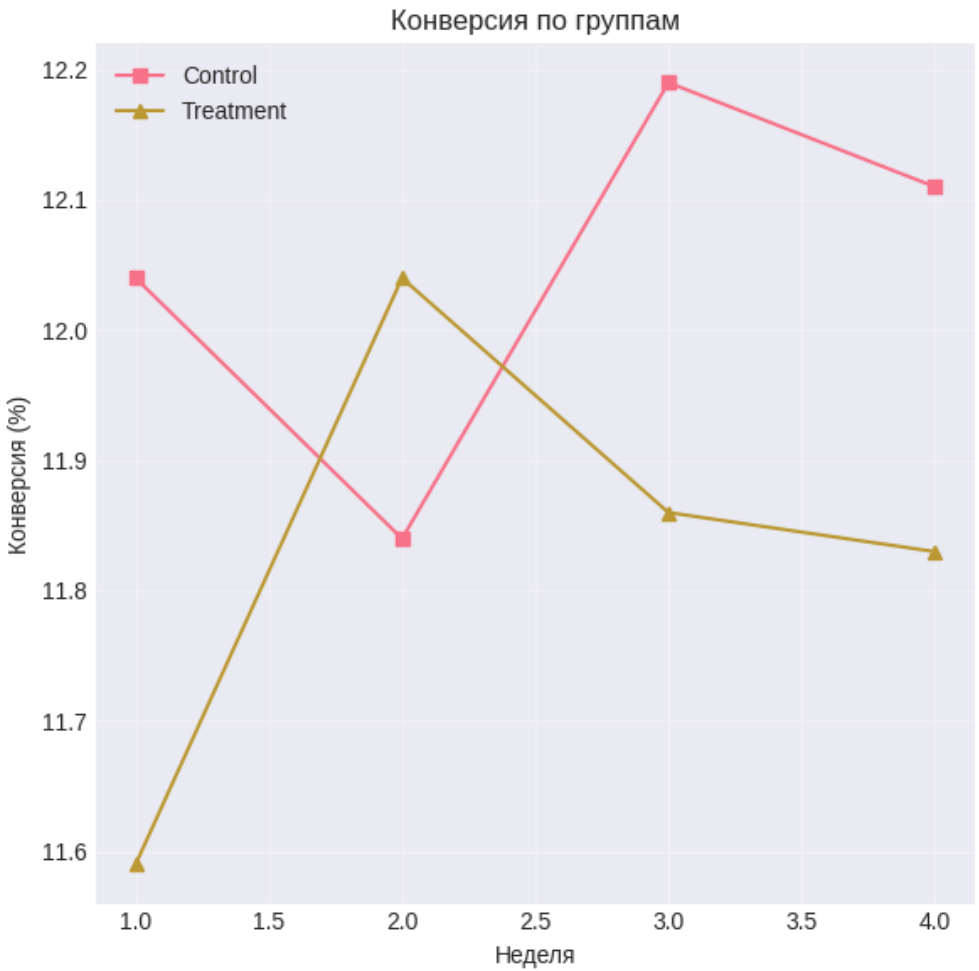
Динамика за весь период:

- **Контрольная группа:** Максимум конверсии (12,8%) на 15-й день, минимумы (11,5-11,6%) на 2-й, 7-й и 10-й дни

- **Экспериментальная группа:** Минимум (11,2%) на 2-й день, ниже контрольного значения

Кумулятивный анализ конверсии по неделям





•Общая конверсия выросла к 3-й неделе, после чего снизилась почти на **0.8%**

- **Среднестатистически** конверсия контрольной группы выше
- **На 2-й неделе** произошло ключевое изменение**:** конверсия контрольной группы снизилась**,** а тритмент выросла и превысила контрольные показатели

Основные выводы EDA

Изменения в обеих группах в течение месяца не превышают **1%**, что указыв ает на **стабильное поведение пользователей** и **отсутствие выраженного Maturity-эффекта**

Конверсия имеет выраженную зависимость от времени - как времени суток, так и дня недели. Пики активности соответствуют естественным ру бежам в распорядке дня пользователей, что позволяет эффективно плани ровать коммуникации или запуск изменений

Выраженная зависимость конверсии от времени является ключевым ограничением при выборе метрики и статистических критериев!

Однородность групп

Для проверки на однородность выборок необходимо проверить ключевые мет
рики, но так как по сути мы знаем только время - его и проверим на однородн
ость

Для этого мы выбрали **тест Колмогорова-Смирнова**, который показал что **p_value = 1.0 >= 0.05** и статистика равна **0.0000 < 0.01** - следовательно, у нас нет причин отклонять нашу **0**-ую гипотезу и различие в выборках мизерное и не и меют бизнес-значения, то есть мы можем считать, что распределения одинако вы

A/A тестирование

На самом деле A/A должно проводиться до самого **A/B** тестировани
я для проверки корректности системы тестирования: проверка пр
авильности сплит-системы, выявление багов системы. Но так как
у нас уже проведен **A/B** тест, то сложно определить правильно ли с
работало распределение, мы не можем точно сказать верно ли бы
ло сделано распределение по группам

*Теперь, когда мы поняли, что наши данные корректны и подходят для проведения **A/B** тестов - мы можем приступить к выбору метрики и па раметров для теста*

Переход к статистическому анализу

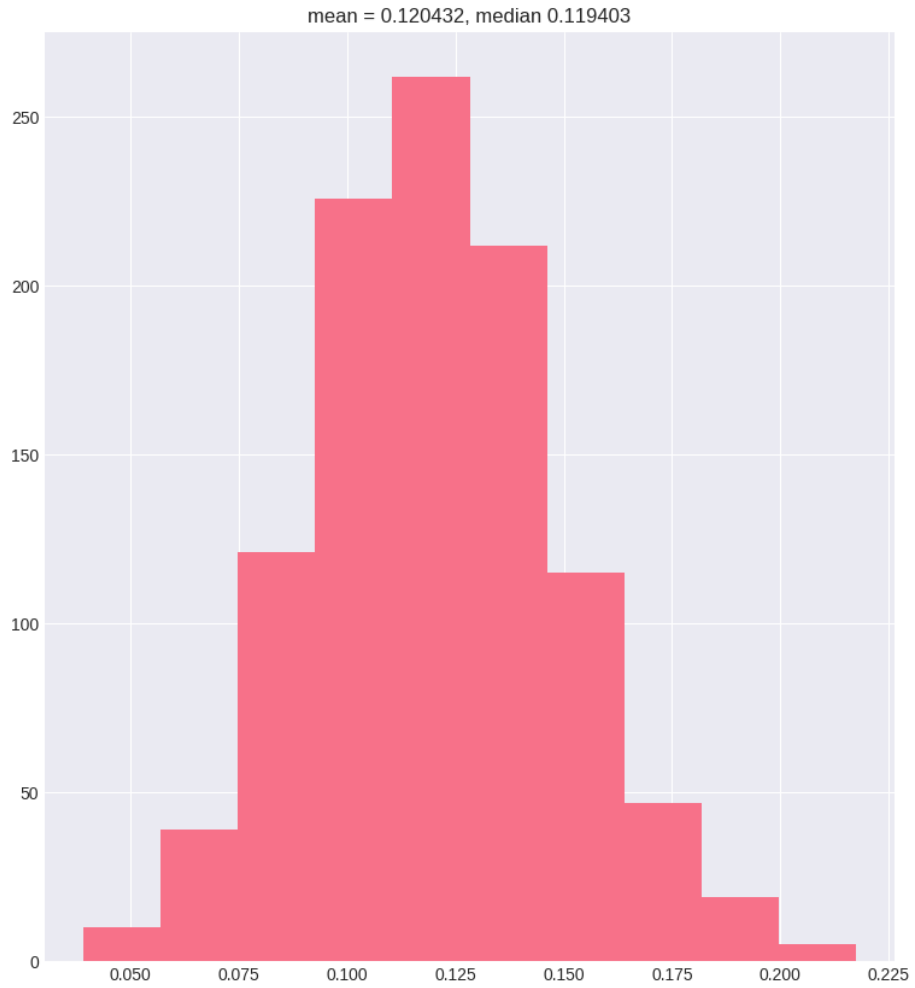
Выбор метрики

Для обеспечения статистической значимости мы будем анализировать не все временные метки, а по агрегиро ванным 30-минутным интервалам.

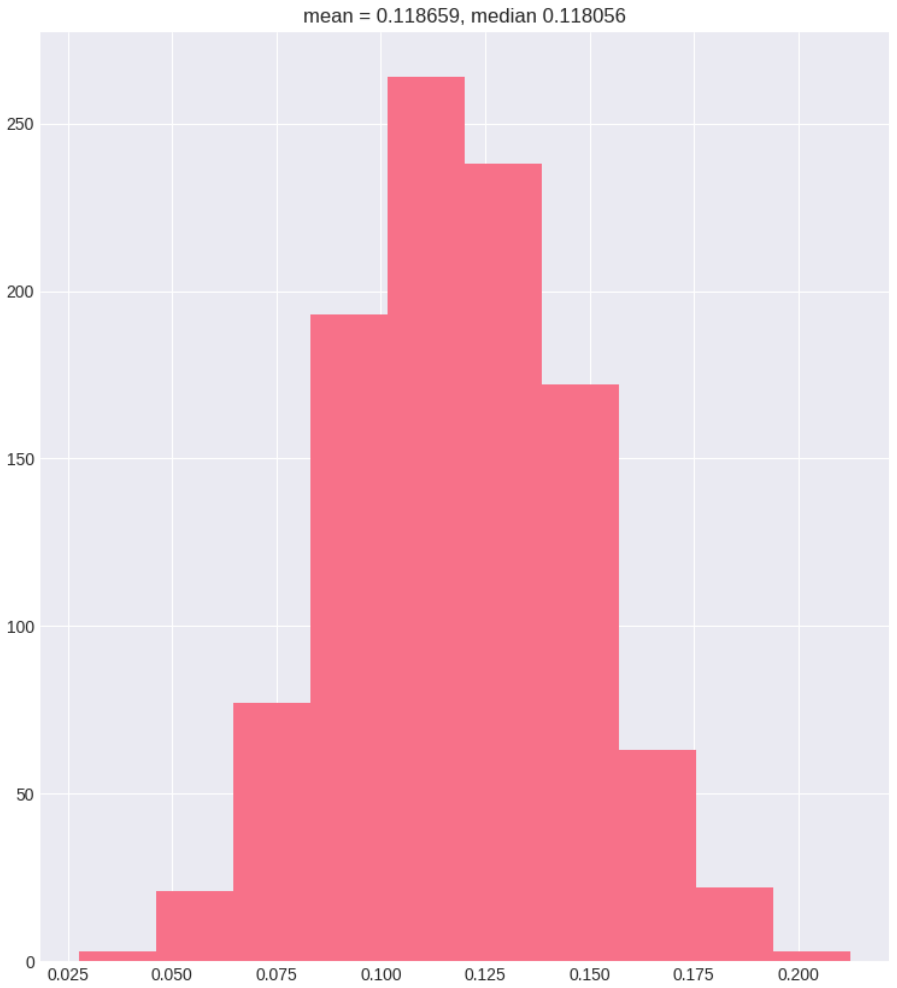
1. В каждом интервале минимум **n = 90** человек. Это число выбрано исходя из требования ЦПТ , которая гара нтирует нормальность распределения выборочных средних при **n ≥ 30.** Это обеспечивает дополнительну ю устойчивость результатов и возможность обнаруживать даже небольшие изменения конверсии.
2. Для каждого временного интервала и для каждой группы мы рассчитываем **CR (Conversion Rate)** - отноше ние просмотров к подпискам. **CR** - это ключевая бизнес-метрика, который показывает какое количество л юдей совершили целевое действие
3. Преимущества похода заключаются в стабильности, так как агрегация по времени снижает влияние време нных аномалий, сравнимость двух групп, временные ряды **CR** позволяют отследить динамику эффективн ости, **CR** напрямую связан с ключевыми бизнес-метриками

Выбор метрики

Контрольная группа



Экспериментальная группа



Определение параметров

При проверке гипотезы мы стремимся принимать точные решения, опираясь не только на статистику, но и на учитывая имеющиеся ресурсы (например, время и деньги)

- По **EDA:** распределения контрольной и тестовой группы схожи.
- При любом выборе выручка, вероятно, не изменится радикально
- Предлагается рассмотреть возможные издержки принятия решения каждой группы

→ При внедрении нового лендинга появляются доп. издержки для создания этого лендинга и его общего внедрения.

Было определено, что **mean_control** (0.120438)> **mean_treatment** (0.118876) **&&** **median_control** (0.11912) > **median_treatment** (0.1182)

Significance level α

Power = 1 - β

MDE

Источник: https://www.google.com/url?q=https%3A%2F%2Fjournals.sagepub.com%2Fdoi%2F10.1177%2F25152459221080396

Type I error: рассмотрели внедрение нового лендинга, хотя более выгодно оставить старый

→**Cost:** создание нового лендинга + риск снижения прибыли (основываясь на средних показателях конверсии)

Type II error: оставили старый лендинг, но выгодней было рассмотреть внедрение нового

→**Cost:**упущенная возможная прибыль в следствии внедрения нового лендинга

Источник: https://www.google.com/url?q=https%3A%2F%2Fjournals.sagepub.com%2Fdoi%2F10.1177%2F25152459221080396

Type I error: рассмотрели внедрение нового лендинга, хотя более выгодно оставить старый

→**Cost:** создание нового лендинга + риск снижения прибыли (основываясь на средних показателях конверсии)

Type II error: оставили старый лендинг, но выгодней было рассмотреть внедрение нового

→**Cost:**упущенная возможная прибыль в следствии внедрения нового лендинга

У ошибка первого рода несёт за собой большие издержки. Теперь посчитаем приближённую «априорную» вероятность

$$P_{H_0} = 0.6; P_{H_1} = 0.4$$

*(стандартный уровень для рассмат риваемой области **0.05**)*

Предлагается применить метод минимизации общего уровня ошибки

$$P_{H_1} * alpha + P_{H_0} * beta \Rightarrow \begin{matrix} \uparrow P_{H_0} \Rightarrow \downarrow alpha \\ \uparrow P_{H_1} \Rightarrow \uparrow alpha \end{matrix} \Rightarrow \begin{matrix} alpha = 0.04 \\ beta = 0.12 \end{matrix}$$

*Для **балансирования ошибки** первого и в торого рода, а также минимизации обще го уровня ошибки, нужно увеличить мощ ность(стандартный уровень для рассма триваемой области **0.2**)*

Источник: https://www.google.com/url?q=https%3A%2F%2Fjournals.sagepub.com%2Fdoi%2F10.1177%2F25152459221080396

$$P_{H_1} * alpha + P_{H_0} * beta \Rightarrow 0.4 * 0.04 + 0.6 * 0.12 = 0.088$$

Получили, что допускаемый уровень совершения ошибки менее 9%!

*(при одинаковых априорных вероятностей и α = **0.05**, β = **0.2** , общий уровень ошибки ~ **0.092**)*

Significance level α = 0.04

Power = 1 – β = 0.88

Источник: https://www.google.com/url?q=https%3A%2F%2Fjournals.sagepub.com%2Fdoi%2F10.1177%2F25152459221080396

$$P_{H_1} * alpha + P_{H_0} * beta \Rightarrow 0.4 * 0.04 + 0.6 * 0.12 = 0.088$$

Получили, что допускаемый уровень совершения ошибки менее 9%!

*(при одинаковых априорных вероятностей и α = **0.05**, β = **0.2** , общий уровень ошибки ~ **0.092**)*

Significance level α = 0.04

Power = 1 – β = 0.88

УТОЧНЕНИЕ:

*Выбор априорных вероятностей, α и β основывается не только на результатах **EDA** и на рассмат риваемой области, но и на **интуиции лица принимающего решение.** Данные оценки часто субъек тивные и определяются лицами, проводящими эксперимент*

Источник: https://www.google.com/url?q=https%3A%2F%2Fjournals.sagepub.com%2Fdoi%2F10.1177%2F25152459221080396

MDE

- Для выбора минимально обнаруживаемого эффекта (**MDE**) для **A/B**-тестирования необходимо согласовать его с бизнес-целями, учитывая практическую значимость (стоит ли вносить изменения?) и ограничения ресурсов

! Пользовательские сайты с высокой посещаемостью: 3-5% MDE обычно считается разумным показателем

- Уровень посещаемости можно определить **mid-to-high**
- Бизнес цель образовательной платформы -это повышение конверсии, следовательно повышения регистраций на курс, а следовательно общей прибыли компании
- Учитывая **EDA**, возможные издержки и средние показатели уровень **MDE**, должен быть низким!
- Средний уровень в отрасли **2-5%.** Так как средние результаты выборок отличаются минимально и важно учитывать риски внедрения изменений , уровень **MDE** устанавливается **1.5%**

MDE = 1.5%

Статистический анализ результатов A/B теста

Для достижения бизнес-цели компания должна определить:
«есть ли вообще смысл что-то делать?»
учитывая возможные риски и издержки. Для этого сформулируем гипотезы:

Глобальные стат. гипотезы:
H_0: В показателях конверсии между двумя группами нет стат. значимой разницы
H_1: В показателях конверсии между двумя группами есть стат. значимая разница

Так как наша ключевая метрика это **CR** с интервальностью **30** минут мы не может точно утверждать, что результаты независимы. То есть данные могут демонстрировать внутри-интервальную корреляцию, тем самым нарушая предположение о независимости и одинаковом распределении

Интервальные гипотезы:

H_0: в к-ом интервале в показателях конверсии между двумя группами нет стат. значимой разницы

H_1: в к-ом интервале в показателях конверсии между двумя группами есть стат. значимая разница

01

Поинтервальное тестирование

*Для определения стат. значимости на каждом временном интервале будет использоваться пропорциональный **Z-test** или **Z-test** для разницы долей (**two-proportion Z-test**).*

1. Для каждого интервала проводим пропорциональный **z-test** и сохраняем **z_score** и **p_value** Вводим счётчик, который будет подсчитывать в скольких тестах **p_value** > **0.04**, то есть в скольких тестах мы приняли H_0

2. Повторяем для всех интервалов

3. Рассматриваем средние и медианные значения **z_score** и **p_value**

ИТОГ:

mean(z_score) = 0.042

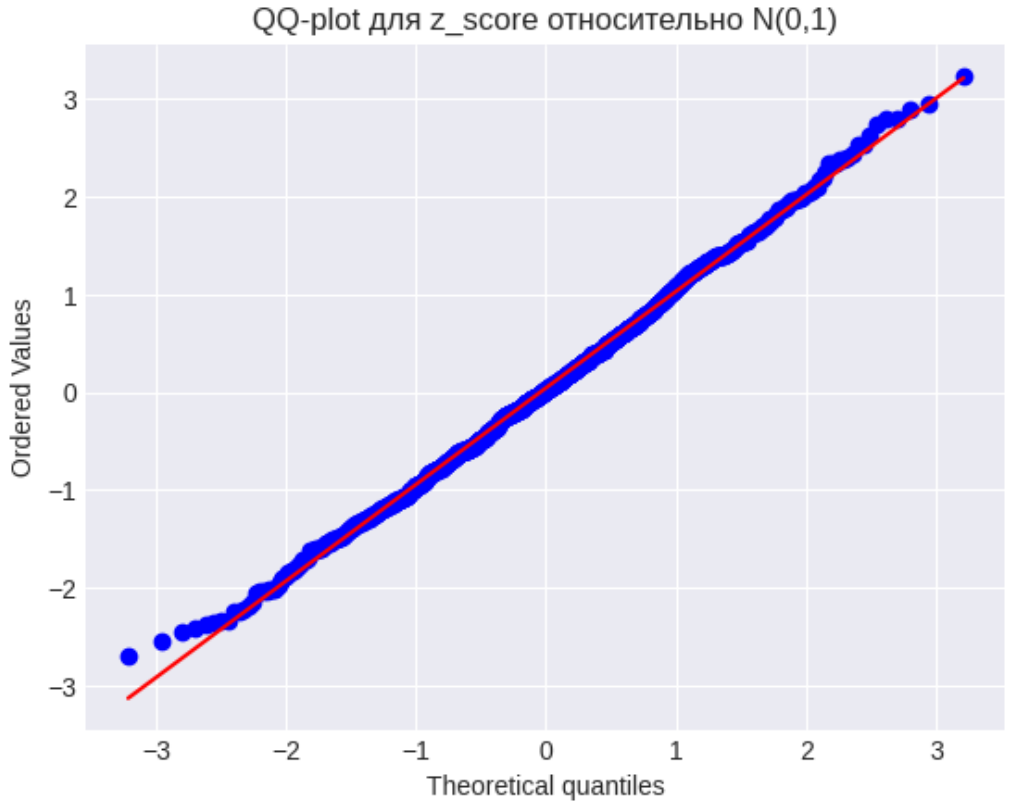
mean(p-value) = 0.504

Получаем, что среднее **p_value** > **0.04.** Значит принимаем нулевую гипотезу, то есть новая версия лендинга не имеет стат. значимое влияние на конверсию

02

Проверка корректности поинтервального тестирования с помощью анализа распределения z-score

- Проверим **z-score** на нормально с помощью теста Колм огорова-Смирнова
- Если **нормальность с центром в нуле** не отвергается и дисперсия стремится к **1**, это подтверждает корректн ость z-теста и **отсутствие систематического сдвига между группами по интервалам.**



ИТОГ:

- имеем **sharipo_pvalue = 0.24 > 0.04, K-S** относительно **N(0,1)** => нет оснований отвергать г ипотезу, что **z_score** распределён аппроксимиров ано **N(0,1)**
- **mean[z_score] -> 0**
- **VAR[z_score] -> 1**

Следовательно можем сделать вывод, что совокупность **тестов принимает первоначальную H_0: новая версия лендинга не имеет стат. значимое в лияние на конверсию**

03

Оценка эффекта с помощью доверительный интервалов

- Идея: рассматриваем эффекты по интервалам как обычную выборку чисел и построить для их среднего доверительный интервал.
- Это даёт ещё один взгляд на глобальную H_0:«в среднем по интервалам эффекта нет»
- Для разности долей **CR, delta_i,** используем нормальную аппроксимацию. Так как выборки достаточно большое (**n=1056**), по ЦПТ разность аппроксимируется, как нормально распределённая случайная величина
- Если ДИ узкий и включает **0**, принимаем H_0, так как H_0 устанавливается, как:
- "отсутствие стат. значимой разницы между контрольной и тритмент группой", те их разница стремится **0**
- Доверительный интервал задаёт диапазон значений параметра, которые «не противоречат» данным на заданном уровне доверия. Поэтому смотрим не только на "попадения **0** в интервал", но и на ширину.
- Имеем: **alpha = 0.04** => строим **96%**-Й ДИ.

ИТОГ:

- ДИ = **[-0.00066, 0.00421]**
- **mean delta = 0.00177 → 0**

Таким образом не имеем достаточно данных, чтобы опровергнуть H_0, а также, если эффект и есть, он настолько мал, что практически неважен.

ИТОГОВЫЙ РЕЗУЛЬТАТ: принимаем H_0

Дополнительные методы анализа: Бета-биномиальное распределение

Бета-биномиальное распределение *— это биномиальное распределение, в котором вероятность успеха в каждом из* ***n*** *испытаний* ***не фиксирована, а выбирается случайным образом из бета-распределения.*** *Оно часто используется в байесовской статистике, эмпирических байесовских методах и классической статистике для выявления избыточной дисперсии в данных с биномиальным распределением.*

- Аппроксимировать свойств Бэта-распределения будем с помощью метода **Монте-Карло,** путём генерации большого числа случайных выборок и эмпирического сравнения
- Вместо аналитического вычисления сложных интегралов или моментов, сэмплируем значения из **posterior** для двух групп, затем считаем долю случаев, где одно превышает другое. Это работает благодаря закону больших чисел: при достаточно больших **N** доля сэмплов сходится к истинной вероятности

(Posterior) ∝ Старые знания **(Prior)** × Новые данные **(Likelihood)**

1.Prior -априорное распределение, это наша гипотеза, что мы определяем до эксперимента:

E[p] = α/(α+β) и уверенность (α+β)

"виртуальные" успехи (α) и неудачи (β)

2. Likelihood (Правдоподобие), что показывают экспериментальные данные. Для каждой возможной конверсии **p** мы спрашиваем: "Насколько вероятно получить такие данные, если истинная конверсия = **p**?"

!Пример:Пусть 180 конверсий из 2000 показов

$$\text{Likelihood}(p) = C(2000, 180) \times p^{180} \times (1-p)^{1820}$$

- **Likelihood** - это "вес" данных для каждой гипотезы.

3. Posterior (Апостериорное распределение), что мы думаем ПОСЛЕ эксперимента

Дополнительные методы анализа: Бета-биномиальное распределение

Для конверсии есть удобное свойство:
Если **Prior** = **Beta**(α, β) и данные = Биномиальные(**k** успехов, **n** испытаний), то:
Posterior = **Beta**(α + **k**, β + **n** - **k**)

Зачем генерировать случайные значения?

Мы не можем аналитически сравнить два **Beta** распределения, но можем:

1. Определить априорные α, β
2. С помощью них определить апостериорные α, β для контрольной и экспериментальной группы
3. Сгенерировать много (**100k**) случайных значений из каждого распределения, вычисляя истинную конверсию контрольной и экспериментальной группы

```
# Posterior для control
alpha_control = alpha_prior + control_conversions
beta_control = beta_prior + control_total - control_conversions # cont

# Posterior для treatment
alpha_treatment = alpha_prior + treatment_conversions
beta_treatment = beta_prior + treatment_total - treatment_conversions
```

```
#абсолютные разности θ_treatment - θ_control для каждой пары сэмплов
diff_samples = p_treatment_samples - p_control_samples

#доля случаев θ_treatment > θ_control = P(θ_t > θ_c | data).
prob_treatment_better = (p_treatment_samples > p_control_samples).mean()
#средняя разность = E[θ_t - θ_c | data].
expected_gain = diff_samples.mean()
```

Дополнительные методы анализа: Бета-биномиальное распределение

ИТОГИ:

- **P(treatment** лучше **Control) = 7.8%**
- Ожидаемый прирост = **-0.171%**
- **96% HDI** для прироста**: [-0.419%, 0.078%]**

- Вероятность**,** что экспериментальная выборка лучше**,** чем контрольная составляет только **7.8 %!**
- Это означает**,** что вероятность ухудшений при принятии гипотезы составит аж **92.2%.**
- Стоит **отклонить альтернативную гипотезу,** и не вводить изменения**,** так как вероятнее все го они **хуже,** чем исходный вариант

Основные бизнес-выводы:

Основной результат: изменения не дают значимого улучшения

01 Риск ухудшения ключевой метрики

Согласно **байесовскому анализу:**

1. Вероятность того, что новая версия лучше контрольной, составляет всего **7.8%**
2. Вероятность ухудшения конверсии при внедрении изменений **около 90%**
3. Ожидаемый прирост конверсии: **-0.171%** (отрицательный), с **96%** доверительным интервалом от **-0.419%** до **+0.078%.**

02 Не внедрять изменения в лендинг в текущем виде

•Инвестиции в развертывание новой версии **не окупятся,** с высокой вероятностью при ведут к падению конверсии**.**

•Если гипотеза о улучшении конверсии остается актуальной - требуется пересмотр диза йна, контента или механик новой версии и проведение нового теста

Основные бизнес-выводы:

Основной результат: изменения не дают значимого улучшения

03 **Не вносить изменения в лендинг, оставить текущую версию.** Предлагается **сохранить бюджет на доработки, повторные тесты.**

Предложения по повторным тестам:

1. Сформировать новые, более конкретные гипотезы на основе качественных исследований
2. Провести анализ данных теста по сегментам
3. Вместо одного изменения тестировать несколько альтернативных вариантов одновременно (A/B/C или многовариантный тест)

Спасибо за внимание!