A/B ТЕСТИРОВАНИЕ НОВОЙ ВЕРСИИ ЛЕНДИНГА

Этот файл – презентация результатов менеджеру, ответственному за продукт.
Проследить за ходом рассуждений, посмотреть на код и другие графики можно в тетрадке Jupyter по адресу:
https://nbviewer.jupyter.org/github/my-secret-account/public_projects/blob/main/12_sbermarket.ipynb (кликабельно)

Задание:

На текущем варианте лендинга пользователь сначала выбирает ритейлера и изучает его каталог товаров. А попытавшись добавить товар в корзину, он может узнать, что из выбранного магазина доставка по его адресу невозможна. В тестовом варианте лендинга пользователи сначала вводят адрес доставки и только потом выбирают один магазин из предложенных. В команде продуктового менеджера надеются на то, что это нововведение поднимет конверсию в добавление товара в корзину без повышения отказов с самого лендинга.

Предыстория:

В проведённом A/B-эксперименте обнаружилось так много проблем, что единственный аналитик в команде потерял душевное равновесие и уехал восстанавливаться. Нужно выручить людей и сделать выводы из проведённого эксперимента. Меня также попросили описать все найденные аномалии, неочевидные или странные зависимости, нестыковки в данных.

План: (кликабельно)

1. НЕДОСТАТКИ В ДАННЫХ

- Тяжелый формат хранения данных
- Инструмент «деления» трафика по группам работает с ошибками
- Неточная запись времени события
- Не хватает события «Закрытие вкладки»
- Некоторые события пропущены
- Записи в логах дублируются

Тяжелый формат хранения данных.

Категориальные столбцы и уникальные идентификаторы, кроме 'retiler_id', записаны в строковом формате, самом невыгодном: они занимают много места в памяти и нуждаются в преобработке. Добавление символов 'UTC' к датам ничем не оправдано, а категориальные переменные стоит кодировать так же, как поле ‘retailer id’ – числами. После обработки использование памяти первой, самой тяжёлой таблицей, сократилось с 126.9 МБ до 7.7 МБ.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 502784 entries, 0 to 502783
Data columns (total 6 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   hit_at         502784 non-null  object
 1   anonymous_id   502784 non-null  object
 2   group          502784 non-null  object
 3   device_type    502784 non-null  object
 4   browser        502784 non-null  object
 5   os             502783 non-null  object
dtypes: object(6)
memory usage: 126.9 MB
```

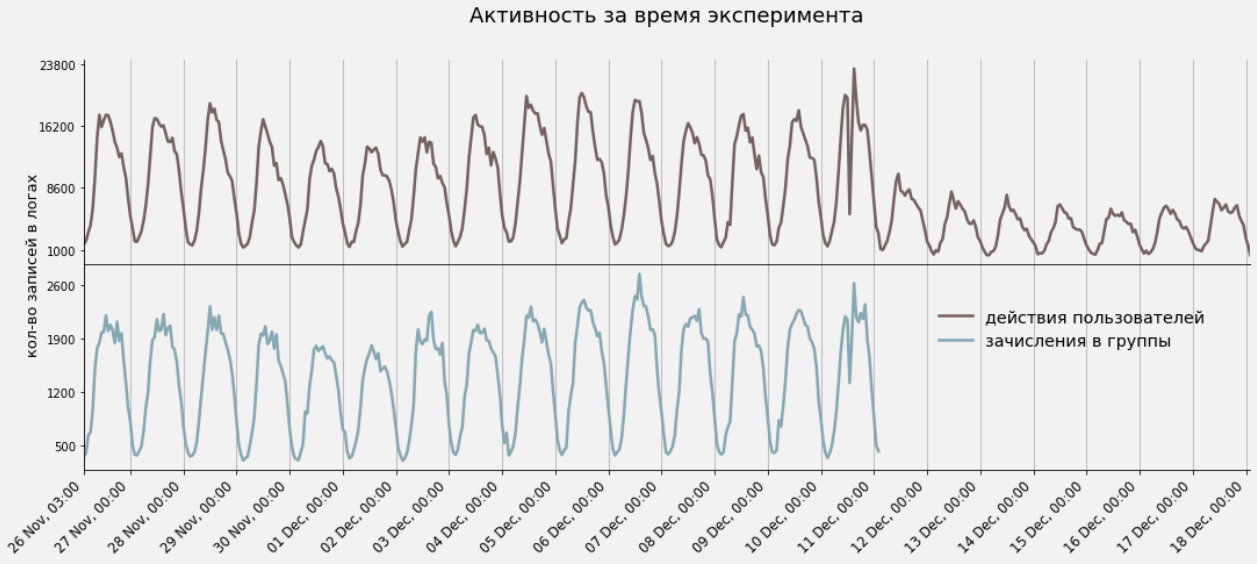
Таблица ‘users’ до предобработки

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 502784 entries, 0 to 502783
Data columns (total 6 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   hit_at         502784 non-null  datetime64[ns]
 1   id             502784 non-null  uint32
 2   group          502784 non-null  uint8
 3   device_type    502784 non-null  uint8
 4   browser        502784 non-null  uint8
 5   os             502784 non-null  uint8
dtypes: datetime64[ns](1), uint32(1), uint8(4)
memory usage: 7.7 MB
```

После предобработки

Инструмент «деления» трафика по группам работает с ошибками.

528 человек меняли группу прямо во время эксперимента. 6604 человек группу не меняли, но появлялись новые записи об определении в группу. В среднем каждый на каждого из них по 2.24 перезаписи.



Как правило, время зачисления в группу совпадает с временем первого появления на сайте.

На графике можно увидеть, что группы пополнялись за счёт входящего трафика. Но оказалось, что зачисление в группу далеко не всегда привязано к попытке пользователя зайти на сайт:

1. 40% уникальных пользователей записаны в какую-то из групп, хотя ни разу ни зашли ни на одну страницу сайта.
2. У 2.5% уникальных пользователей разница во времени (задержка) между определением в группу и первым действием превышает 15 минут. Начиная с этого порога, задержка распределена случайным образом – принимает значения от нескольких часов до нескольких недель. Иногда зачисление в группу сильно запаздывает за первым действием, как в примере ниже.
3. Перезаписи то в одну, то в другую группу в 9 раз чаще происходили с теми пользователями, у которых задержка между первым действием и определением в группу превышает 4 минуты.

4. Перезаписи в 4 раза чаще происходили с теми, кто ни разу не вводил свой адрес, но добавлял товары в корзину. Но тот, кто впервые оказался на сайте, не может добавить товар, не введя адрес. Единственный вариант – когда старый пользователь передаёт куки, и в них уже содержится введённый ранее адрес. В итоге, проблема с перезаписью в разные группы связана, во-первых, с большими задержками до и после определения в группу и, во-вторых, со старыми пользователями.

До определения в группу прошло две недели.

	timestamp	id	event	group
110237	2020-11-28 15:11:54.982000	050674f7-f90a-40db-9193-6d49365cf478	Каталог	0
110238	2020-12-10 12:09:44.114000	050674f7-f90a-40db-9193-6d49365cf478	Определён в группу	0
110239	2020-12-11 10:55:15.403000	050674f7-f90a-40db-9193-6d49365cf478	Лендинг	0
110240	2020-12-11 10:55:45.543000	050674f7-f90a-40db-9193-6d49365cf478	Каталог	0
110241	2020-12-11 10:55:55.729000	050674f7-f90a-40db-9193-6d49365cf478	Магазин выбран	0
110242	2020-12-11 10:55:59.923000	050674f7-f90a-40db-9193-6d49365cf478	Каталог	0
110243	2020-12-11 10:56:52.584000	050674f7-f90a-40db-9193-6d49365cf478	Магазин выбран	0

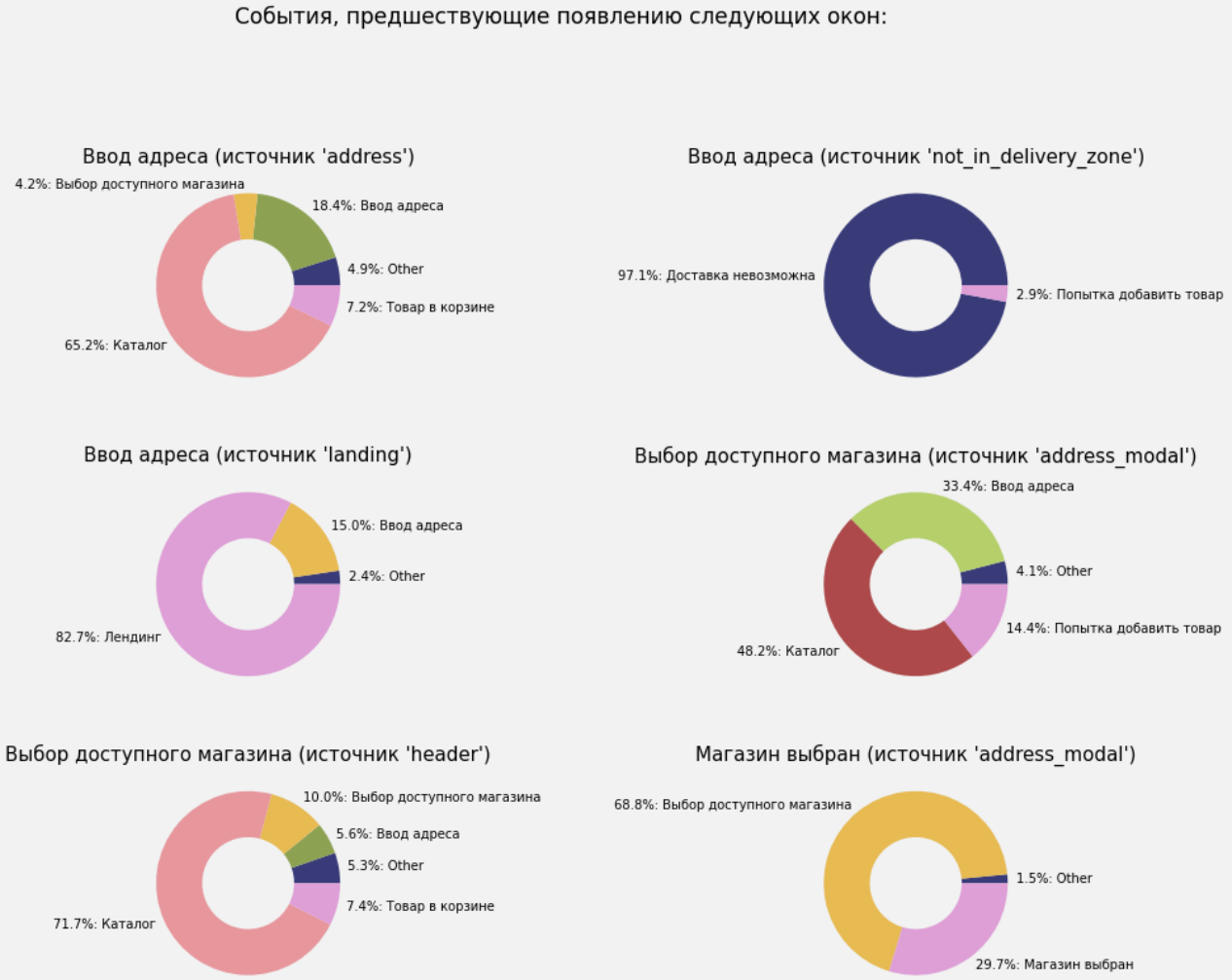
Создается впечатление, что одновременно работают два механизма привлечения трафика в эксперимент. Один скрипт привлекает в группы текущих посетителей сайта, а другой скрипт итерируется по базе известных пользователей. Иногда эти скрипты накладываются на одних и тех же людей, в результате один человек может быть зачислен в группу несколько раз.

Неточная запись времени события.

Время регистрации события не всегда совпадает со временем, когда оно произошло. Из-за этого события могут выстраиваться в неправильном порядке. Судя по графику справа, долю перепутанных во времени записей можно оценить от 1 до 5 процентов, – это величина категории ‘Other’. Категория ‘Other’ появляется на графиках постоянно, для какого бы значения ‘source’ мы не искали предыдущие события. В неё входят те события, которые никак не могут предшествовать вызову модального окна с указанным источником. Например, попытка добавить товар никак не может вызвать окно для ввода адреса из источника 'not_in_delivery_zone' – но в данных такие случаи присутствуют (это видно на правом верхнем pie chart’е).

Не хватает события «Закрытие вкладки».

– Для построения таких графиков, как справа, для разделения пользовательского опыта для сессии и для многих других аналитических задач приходится выбирать произвольный и негибкий порог, например, в 20 минут, только для того, чтобы компенсировать отсутствие этого события в логах.



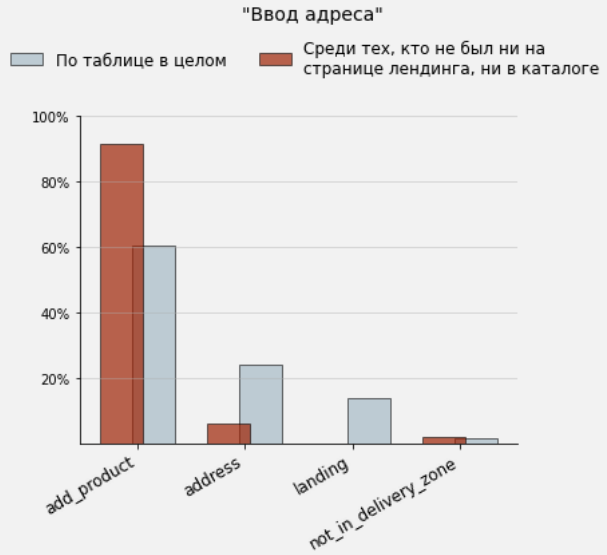
Некоторые события пропущены.

Эту проблему удалось обнаружить на пользователях, которые, если верить логам, не посещали ни лендинг, ни каталог. Дело в том, что кроме них, у нас есть 4 модальных окна, которые не открываются по прямой ссылке и два действия, которые недоступны новым пользователям, пропустившим предыдущие окна. Отсюда два варианта: либо в логах пропущенные записи, либо – старые пользователи.

Из них только лишь 7.6% доходили до добавления товара в корзину, – это не похоже на картину по старым пользователям. 87% из них вводили адрес, из них примерно 90% – через модальное окно, которое открывалось из источника ‘add_product’ (после попытки добавить товар в корзину). Остальные 10% – с помощью кнопки, которая лежит в шапке каталога. И те, и другие пользовались каталогом, запись о котором пропущена.

Пользователь явно обновлял страницу каталога несколько раз. Но где запись?

	timestamp	id	event	group	source
102724	2020-12-10 10:23:29.450000	240e260b-69b4-4585-b72c-c890ae89fc51	Выбор доступного магазина	0	
102725	2020-12-10 10:23:30.427000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102726	2020-12-10 10:23:30.537000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102727	2020-12-10 10:23:45.804000	240e260b-69b4-4585-b72c-c890ae89fc51	Выбор доступного магазина	0	
102728	2020-12-10 10:24:35.546000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102729	2020-12-10 10:24:35.558000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102730	2020-12-10 10:24:35.612000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102731	2020-12-10 10:25:09.449000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102732	2020-12-10 10:25:09.505000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102733	2020-12-10 10:25:36.829000	240e260b-69b4-4585-b72c-c890ae89fc51	Выбор доступного магазина	0	
102734	2020-12-10 10:25:37.086000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102735	2020-12-10 10:25:37.156000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header
102736	2020-12-10 10:25:39.300000	240e260b-69b4-4585-b72c-c890ae89fc51	Магазин выбран	0	header



Записи в логах дублируются.

Иногда дублирование записи предусмотрено и сделано намеренно – например, если по введённому адресу доставка невозможна, тогда регистрируется новое событие «Ввод адреса» с источником (‘source’) – 'not_in_delivery_zone'. Но в большинстве случаев последовательные, или дублированные записи ни о чём не говорят и никак не объясняются. К примеру, пятая часть от всех зарегистрированных событий «Магазин выбран» – это записи, следующие одна за другой с промежутком меньше 133 миллисекунд. Может быть, существуют обстоятельства, при которых компьютер пользователя отправляет несколько одинаковых запросов к серверу одновременно.

2. НЕДОСТАТКИ В ДИЗАЙНЕ ЭКСПЕРИМЕНТА

- В эксперименте должны были участвовать только новые пользователи
- Несбалансированные группы, нет отбора трафика
- Пользователям оставлена возможность не использовать тестируемый интерфейс
- Параллельное тестирование нескольких изменений
- Эксперимент не подготовлен
    I. Две недели и один день: один день был лишним
    II. Недостаточно документации
    III. Не продумано заранее, какие данные нужны, а какие нет
    IV. Не предусмотрено A/A – тестирование
    V. Параллельное тестирование нескольких изменений

В эксперименте должны были участвовать только новые пользователи.

Страница лендинга у групп отличается. У страницы, которую мы показываем контрольной группе, на первом плане магазины-партнёры и фраза "Выберите магазин и начинайте делать покупки" – она пытается вовлечь пользователя брендами. У тестовой группы на первом плане поисковая строка и фраза "Введите адрес доставки и мы покажем...". Эта страница продаёт услугу по доставке. Задача начальной страницы в том, чтобы лаконично объяснить незнакомому человеку, в чём ценность нашего продукта, а тут в эксперименте обнаружилась неопределённо большая группа таких людей, которые уже сформировали мнение о нём – они обратят мало внимания на заголовки и на промо, даже если попадут на лендинг.

Проблема не только в разнице между первым и вторым впечатлением, но и в плане отслеживания, и в метриках. Конверсия старых пользователей в добавление товаров по определению выше, чем у тех, кто впервые попал на сайт, и она совершенно не зависит от того, какую версию лендинга мы им показываем. А воронка, которую они проходят, даже если начинают сессию со страницы лендинга, совершенно другая – они перескакивают через этапы.

Отследить, насколько равномерно они распределены по экспериментальным группам, или чем закончились их предыдущие сессии, тоже не получится – таких данных нет. Хотя если верно, что существует скрипт, который итерируется по исторической базе данных, то можно было бы избежать многих сложностей, просто добавив переменную ‘lifetime’ к таблице ‘users’.

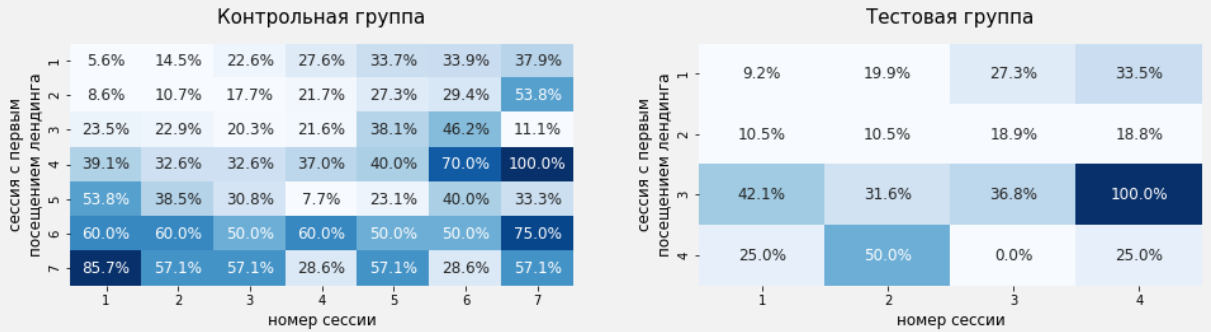
Несбалансированные группы, нет отбора трафика.

Размеры контрольной и тестовой группы соотносятся как 9:1. С этим нет особых проблем до тех пор, пока группы однородны. Если в них обнаруживается подгруппа, и подгруппа отличается по значению целевой метрики, то любой сдвиг в распределении подгруппы приносит несоразмерные изменения в метрику наименьшей по размеру группы. Без отбора трафика, таких подгрупп сколько угодно много, но становится известно о них лишь по косвенным признакам. Если изучить таблицу, которая ниже, можно заметить подгруппу «продвинутых» пользователей – тех, кто не нуждается в лендинге и может обходиться без него больше четырёх сессий.

Из-за несбалансированности групп:

1. Распределение подгруппы трудно отследить. В тестовой группе не хватает данных – к четвёртой когорте остаются лишь 4 пользователя, притом это нормально – размер наименьшей из экспериментальных групп обычно планируют таким, чтобы хватало на тест целевой метрики. Для исследования подмножеств размера уже недостаточно. Кроме того, один процент данных – это много в масштабе тестовой группы, но мало в масштабе контрольной. Выход – пересчитывать абсолютные величины в относительные. Но из-за спешки, неопытности или неуверенного владения языками программирования здесь можно допустить много ошибок.
2. Удалив данные, распределение подгруппы можно непредсказуемым образом перекосить. Если до этого оно было случайным, как и отбор в экспериментальные группы, то теперь будет зависеть от признаков, которые мы смогли выделить. Любой перекос будет иметь тем более тяжелые последствия, чем сильнее несбалансированы группы. Это – известный парадокс Симпсона. Когда при тестировании вакцины от болезни, который женщины подвержены меньше мужчин, в группе плацебо несоразмерно много женщин, данные об эффективности вакцины становятся заниженными.
2. Если данные не удалить, то результаты будут зашумлены такими значениями метрик, которые слабо или неочевидно связаны с тестируемым изменением. Продвинутые пользователи отлично конвертируются в добавление товара независимо от сценария, который прописан в лендинге.

Сессий с успешными попытками добавить товар в корзину:



Отбором трафика можно было избежать и старых пользователей, и тех, кто начал знакомство с сайтом не с лендинга, и предотвратить другие подгруппы. Таргетировать изменения, то есть сужать аудиторию, на которую они направлены (как минимум на время эксперимента) – это в любом случае хорошая практика.

Пользователям оставлена возможность не использовать тестируемый интерфейс.

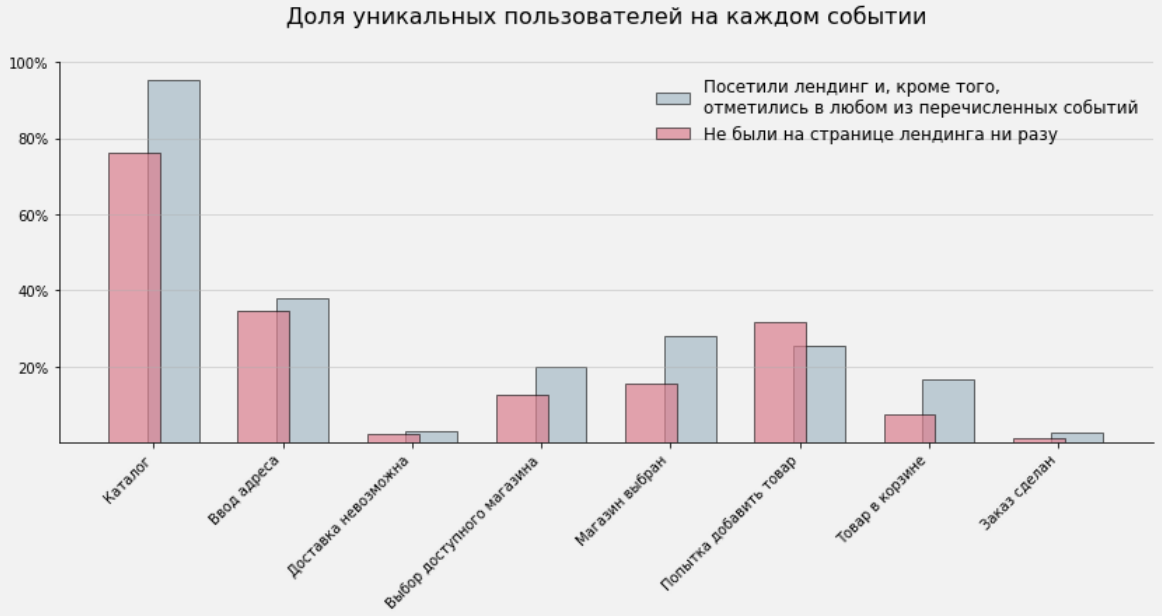
Из 19 000 человек, которые успешно добавляли товары в корзину, 5540 ни разу не заходили на лендинг. А среди тех, кто не пополнял корзину, на начальной странице не были 67 920 человек (четверть всех уникальных пользователей).

Идея тестируемого интерфейса в том, чтобы избежать ситуаций, когда пользователь, находясь уже в одном шаге от покупки, узнаёт, что не сможет купить выбранный им товар. Но спроектирована новая версия лендинга так, что такая возможность остаётся. Несмотря на то, что ни один переход с новой версии страницы не ведёт в каталог, попасть туда всё равно можно, например, по ссылке. 25% прямого (direct) трафика на сайт – обыкновенная вещь. Если не по прямой ссылке, то после входа в личный кабинет по номеру телефона (к слову, а почему это событие не отслеживается?). Так или иначе, обе версии сайта работают по одной и той же механике с теми, кто пропустил лендинг.

Но если не очистить тестовую группу от таких пользователей, то метрика конверсии в ней просядет (это видно на графике), притом незаслуженно. Ведь в ситуацию с «кривым сценарием» не могут попасть те пользователи из тестовой группы, навигация которых началась с лендинга. Напротив, в контрольной группе метрика просядет «заслуженно», но чтобы предотвратить парадокс Симпсона, о котором шла речь на предыдущей странице, таких пользователей стоит исключить из обеих групп.

	timestamp	id	event	group	session	source
1073526	2020-12-04 08:52:27.857000	01ca9617-2b62-4488-921b-2de58f54322e	Лендинг	1	1	
1073527	2020-12-04 08:52:35.177000	01ca9617-2b62-4488-921b-2de58f54322e	Каталог	1	1	
1073528	2020-12-04 08:52:45.789000	01ca9617-2b62-4488-921b-2de58f54322e	Ввод адреса	1	1	address
1073529	2020-12-04 08:52:57.005000	01ca9617-2b62-4488-921b-2de58f54322e	Каталог	1	1	
1073530	2020-12-04 08:53:02.068000	01ca9617-2b62-4488-921b-2de58f54322e	Магазин выбран	1	1	
1073531	2020-12-04 08:53:06.271000	01ca9617-2b62-4488-921b-2de58f54322e	Каталог	1	1	
1073532	2020-12-04 08:53:39.978000	01ca9617-2b62-4488-921b-2de58f54322e	Магазин выбран	1	1	
1073533	2020-12-04 08:53:43.490000	01ca9617-2b62-4488-921b-2de58f54322e	Каталог	1	1	
1073534	2020-12-04 08:55:55.766000	01ca9617-2b62-4488-921b-2de58f54322e	Магазин выбран	1	1	
1073535	2020-12-04 08:55:59.989000	01ca9617-2b62-4488-921b-2de58f54322e	Каталог	1	1	
1073536	2020-12-04 09:05:08.192000	01ca9617-2b62-4488-921b-2de58f54322e	Магазин выбран	1	1	
1073537	2020-12-04 09:05:11.614000	01ca9617-2b62-4488-921b-2de58f54322e	Каталог	1	1	
1073538	2020-12-04 09:05:17.032000	01ca9617-2b62-4488-921b-2de58f54322e	Попытка добавить товар	1	1	
1073539	2020-12-04 09:05:18.602000	01ca9617-2b62-4488-921b-2de58f54322e	Товар в корзине	1	1	

Пользователь был на тестируемом лендинге, но его путь до пополнения корзины – это типичный путь для контрольной, а не тестовой группы



По экспериментальным группам я бы распределял тех пользователей, которые, судя по нашим данным, заходят на сайт впервые. Притом не сразу, а когда они начинают взаимодействовать с лендингом – так можно было избежать вариативности в сценариях сессий. При таком подходе не получится подсчитать метрику Bounce, но под неё вовсе не обязательно собирать данные – её хорошо считает Яндекс.Метрика.

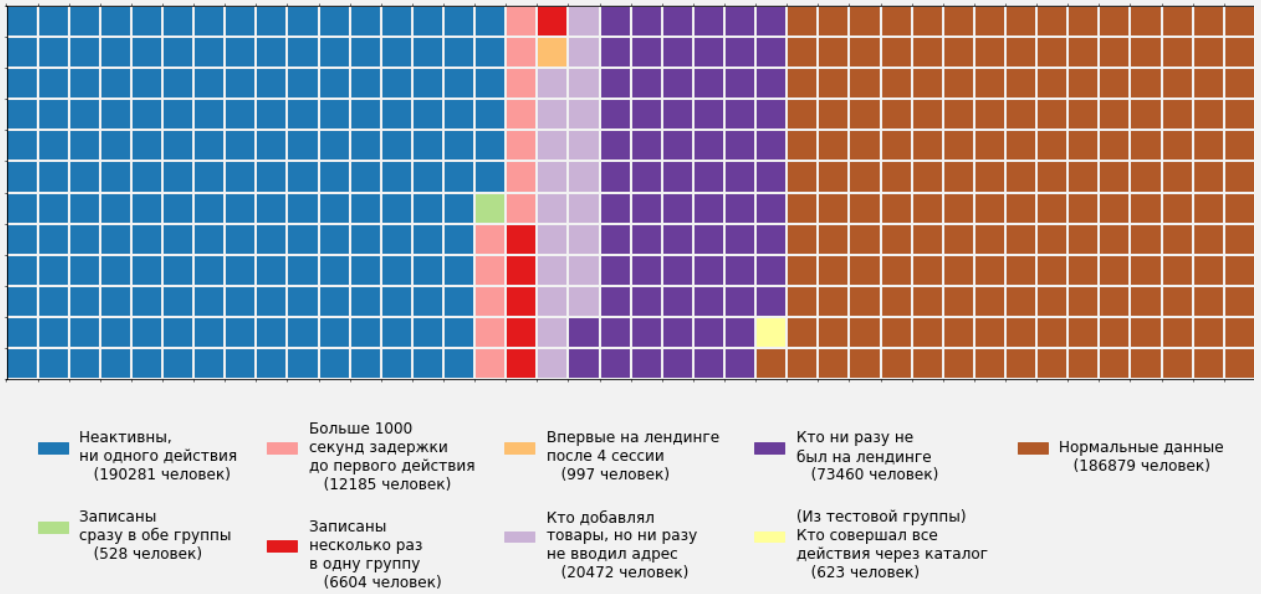
	timestamp	id	event	group	session	source
1404022	2020-11-29 05:46:31.159000	2a57cde4-2b33-49ce-947d-c653823df63b	Лендинг	1	1	
1404023	2020-11-29 05:46:31.887000	2a57cde4-2b33-49ce-947d-c653823df63b	Ввод адреса	1	1	landing
1404024	2020-11-29 06:12:31.281000	2a57cde4-2b33-49ce-947d-c653823df63b	Выбор доступного магазина	1	1	
1404025	2020-11-29 06:12:33.147000	2a57cde4-2b33-49ce-947d-c653823df63b	Магазин выбран	1	1	
1404026	2020-11-29 06:12:33.189000	2a57cde4-2b33-49ce-947d-c653823df63b	Магазин выбран	1	1	
1404027	2020-11-29 06:12:36.592000	2a57cde4-2b33-49ce-947d-c653823df63b	Каталог	1	1	
1404028	2020-11-29 06:13:01.021000	2a57cde4-2b33-49ce-947d-c653823df63b	Попытка добавить товар	1	1	
1404029	2020-11-29 06:13:07.787000	2a57cde4-2b33-49ce-947d-c653823df63b	Каталог	1	1	
1404030	2020-11-29 06:13:21.451000	2a57cde4-2b33-49ce-947d-c653823df63b	Попытка добавить товар	1	1	

Пример того, как может не хватать документации.
Почему пользователь не смог добавить товар?

Эксперимент не подготовлен.

Не до конца продуманы, как мне показалось, ещё несколько моментов:

1. Несколько неудачно, что будних дней на один больше. Всех дней недели по два, а четверга – три. Профиль поведения на выходных и на буднях зачастую отличается, в том числе в нашем кейсе – доставку удобно заказывать именно в те дни, когда ты находишься дома. На доставке сфокусирована тестовая версия лендинга. Контрольная – на тех магазинах-партнёрах, куда, скажем, большая часть населения привыкла ходить после работы в будние дни. Неочевидно, зачем сделано так, если эксперимент можно было не в ущерб себе закончить ровно после двух недель.
2. Не хватило документации, в которой описывается порядок ведения логов (особенно в контексте разницы между группами). Не хватило описаний тех принципов, на которых основано распределение по экспериментальным группам. Если команде известны такие проблемы, как временные задержки до записи события или как дублирование событий в логах, то о них следовало упомянуть. Чтобы понять, что означают категории в переменных «source», пришлось проводить исследование внутри исследования, а в результате вопросов стало ещё больше.



3. Гипотеза, которая тестируется, сформулирована не «по-аналитически». Если бы в команде описали аудиторию, на которую рассчитано нововведение, описали метрику, которая должна вырасти, привели текущие данные и ожидаемое улучшение, то многих ошибок и несостыковок удалось бы избежать на стадии подготовки эксперимента.
4. Запуская эксперимент, могли почувствовать, что он недостаточно спроектирован, что в нём могут обнаружится проблемы, и целевую метрику из данных придётся извлекать хирургическим путём. Если так, то с учётом того, как много трафика запущено в контрольную группу, можно было, ничего не теряя, создать вторую контрольную группу и дать аналитику возможность «сверять часы» на A/A – тесте.
5. Строго говоря, пропустившие страницу лендинга пользователи всё-таки имеют дело с разными версиями сайта. У контрольной группы окно, сообщающее о недоступности доставки, перекрывает предыдущие окна, содержит строчку текста и большую кнопку, нажав на которую, можно вернуться обратно. У тестовой группы окно ввода адреса в этом случае остаётся прежним, со своей картой и полем для ввода адреса. Единственное отличие – появляется надпись красного цвета. Это изменение можно протестировать отдельно, но не параллельно с тестированием двух версий лендинга, ведь целевая метрика не является независимой от того, как выглядит и как работает это окно.

3. A/B ТЕСТИРОВАНИЕ И ЧТО ДЕЛАТЬ

Стандартные способы подсчета метрик: $\text{Bounce} = \left(1 - \frac{\text{переходы с лендинга}}{\text{показы лендинга}}\right) \cdot 100\%$ $\text{Целевая конверсия} = \frac{\text{пополнения корзины}}{\text{переходы с лендинга}} \cdot 100\%$

Формулы нуждаются в поправках на нестандартные ситуации в данных, иначе статус многих пользователей мы оценим неверно.

BOUNCE RATE

Контрольная группа.

Те из них , кто не совершал действий, помимо просмотра лендинга

Все пользователи, **первая сессия** которых содержит посещение лендинга

Тестовая группа

Те из них, кто не совершал действий, помимо просмотра лендинга

Все пользователи, **первая сессия** которых содержит посещение лендинга. **Если** есть вызов модальных окон, то хотя бы одно – из источника ‘landing’

КОНВЕРСИЯ ИЗ АКТИВНОСТИ В ДОБАВЛЕНИЕ ТОВАРА В КОРЗИНУ

Контрольная группа

Те из них, кто хотя бы один раз добавил товар в корзину

Все, кто совершил хотя бы одно действие, помимо просмотра лендинга

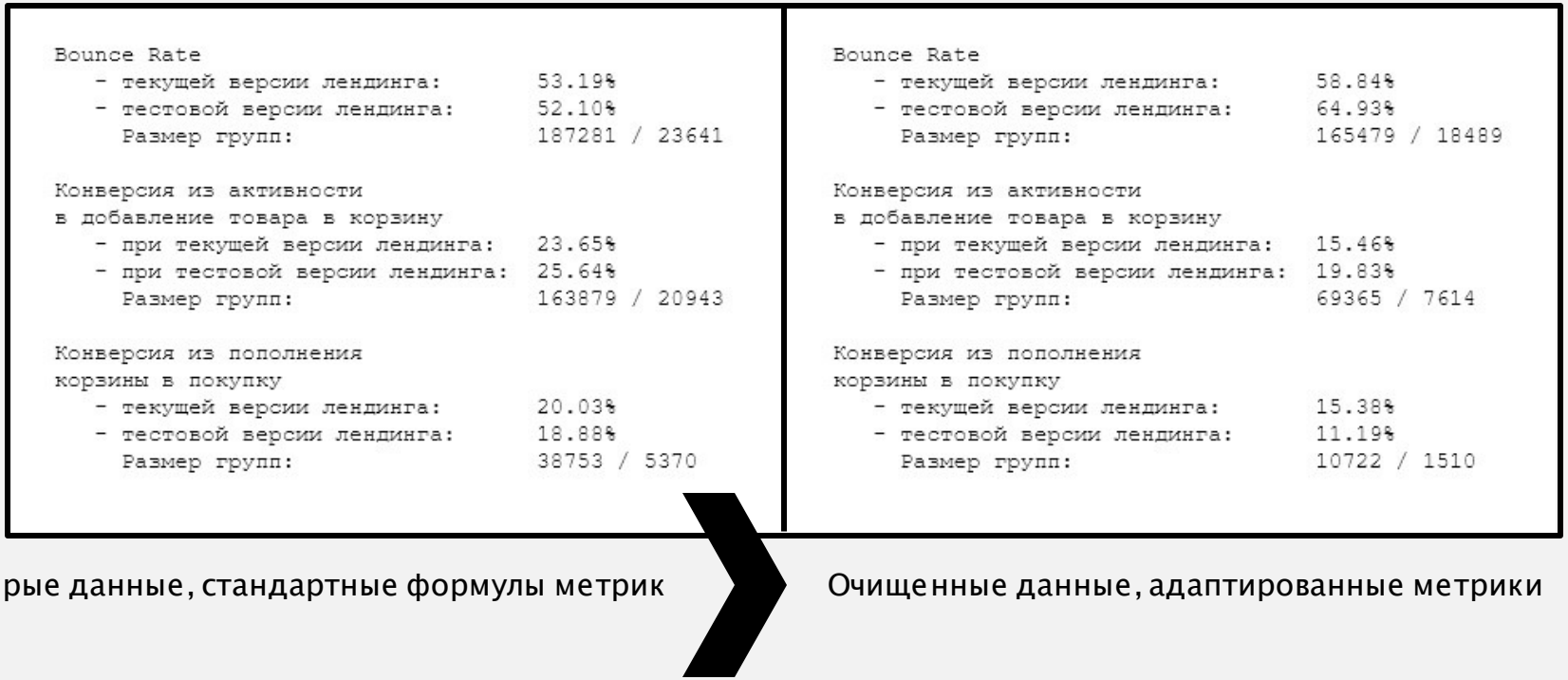
Тестовая группа

Те из них, кто хотя бы раз добавил товар в корзину **и** выбрал магазин в модальном окне, вызванном из источника ‘landing’

Все, кто хотя бы раз вызвал **окно ввода адреса** из источника ‘landing’

Мы не сможем принять решение, не посмотрев на конверсию в покупки и на то, отличается ли выручка с пользователей в разрезе экспериментальных групп. Ведь цель, в конечном итоге, в том, чтобы пользователи совершали покупки, и чтобы с их покупок была прибыль.

Конверсия в покупку считается, как и другие конверсии, по пользователям, а не событиям. В знаменателе – «успехи» предыдущей метрики. В числителе – те, о ком есть хотя бы одна запись «Заказ оставлен».



Bounce Rate	
- текущей версии лендинга:	53.19%
- тестовой версии лендинга:	52.10%
Размер групп:	187281 / 23641
Конверсия из активности в добавление товара в корзину	
- при текущей версии лендинга:	23.65%
- при тестовой версии лендинга:	25.64%
Размер групп:	163879 / 20943
Конверсия из пополнения корзины в покупку	
- текущей версии лендинга:	20.03%
- тестовой версии лендинга:	18.88%
Размер групп:	38753 / 5370

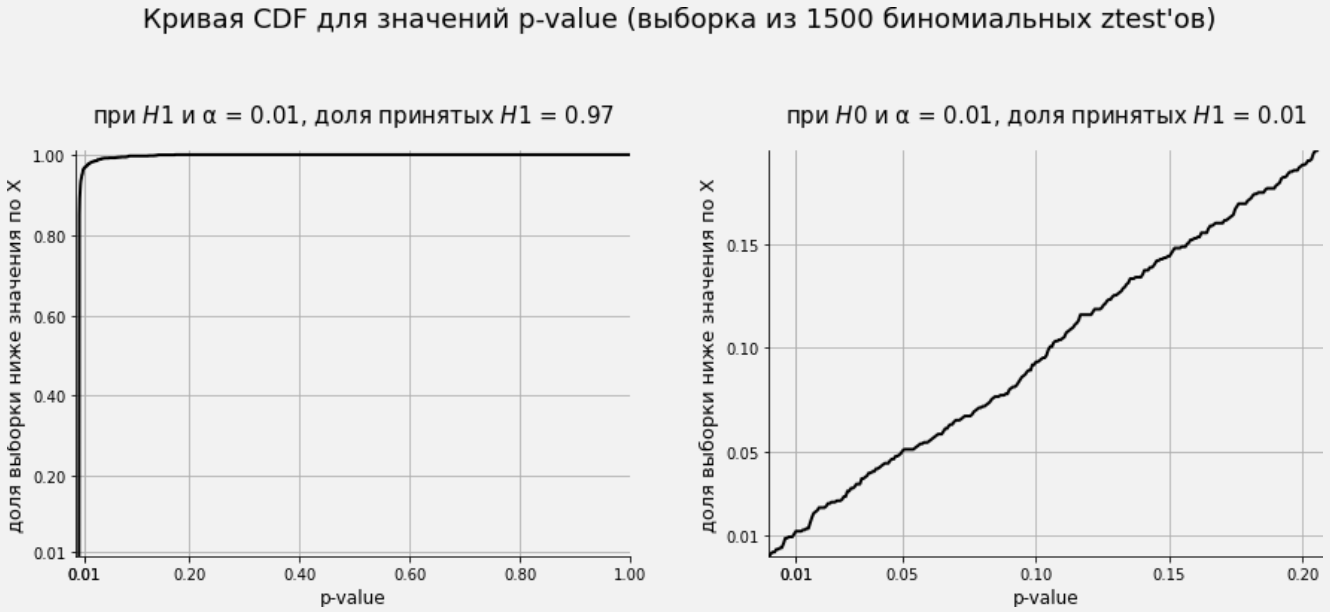
Bounce Rate	
- текущей версии лендинга:	58.84%
- тестовой версии лендинга:	64.93%
Размер групп:	165479 / 18489
Конверсия из активности в добавление товара в корзину	
- при текущей версии лендинга:	15.46%
- при тестовой версии лендинга:	19.83%
Размер групп:	69365 / 7614
Конверсия из пополнения корзины в покупку	
- текущей версии лендинга:	15.38%
- тестовой версии лендинга:	11.19%
Размер групп:	10722 / 1510

Сырые данные, стандартные формулы метрик

Очищенные данные, адаптированные метрики

На всех метриках замена лендинга сказалась очень сильно. Сильнее, чем это казалось по зашумлённым данным, а Bounce Rate и вовсе развернулся в обратную сторону. Это можно понять – когда с порога просят твой адрес, то уже не так привлекательно, как было до этого. Можно найти объяснение и тому, что конверсия в покупку в тестовой группе выше: в тестовой группе пользователи вводят адрес, не зная, что их ожидает дальше. В дальнейшем многие из них будут пополнять корзину от скуки. В то же время, в контрольной группе адрес вводят те, кто уже присмотрел что-то для себя и попытался добавить товар в корзину. Притом большинство из них, введя адрес, сразу получают подтверждение: «всё в порядке, мы привезём продукты по тому самому адресу, который вы ввели». Неудивительно, если некоторых это обнадёживает и подталкивает к покупке.

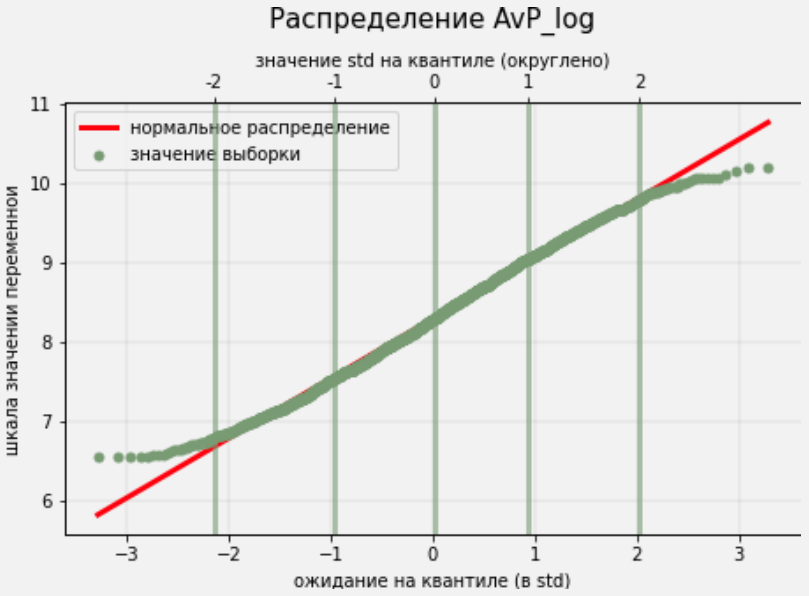
Вряд ли будут проблемы с принятием различий при таких значительных uplift'ах и больших выборках.



Конверсия не зависит от числа показов лендинга. Будем считать исход для одного пользователя независимой случайной величиной, распределённой по Бернулли. В этом случае лучше всего для симуляции данных подходит биномиальное распределение, а в качестве критерия – биномиальный z-test.

На рисунке слева – показатели этого теста при оценке различий между группами в конверсии из корзины в покупку. Различия именно в этой метрике принять тяжелее всего (по сравнению с остальными двумя метриками) и, тем ни менее, у ztest'а сохраняется отличная чувствительность при низком пороге значимости.

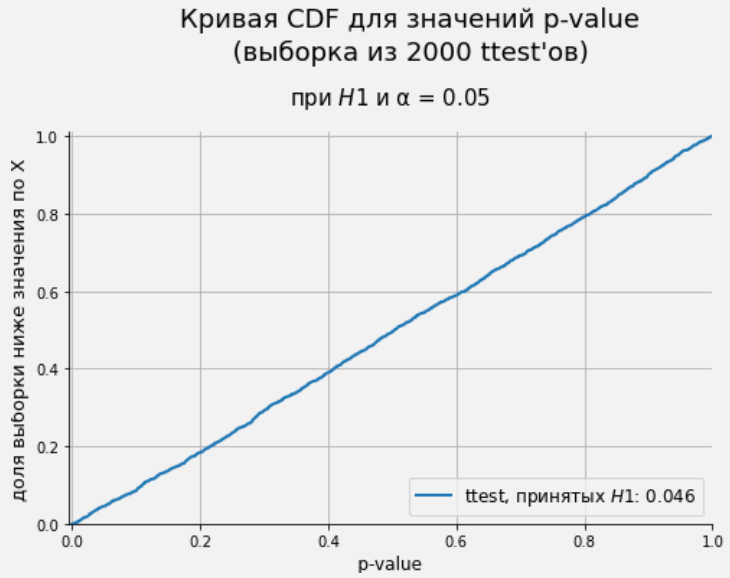
Насчёт среднего чека. Выяснилось, что при взятии логарифма распределение переменной становится близко к нормальному. Ассиметрии нет, дисперсия в группах равная – в итоге, вопрос о выборе оптимального статистического метода не стоит особо остро (подойдёт почти любой).

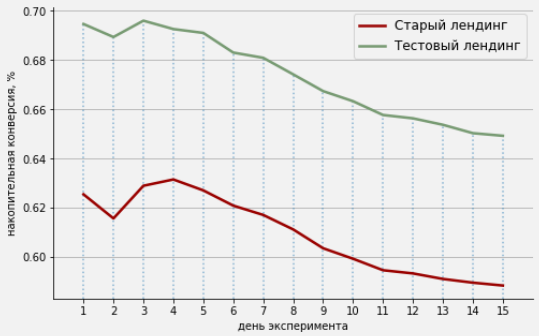


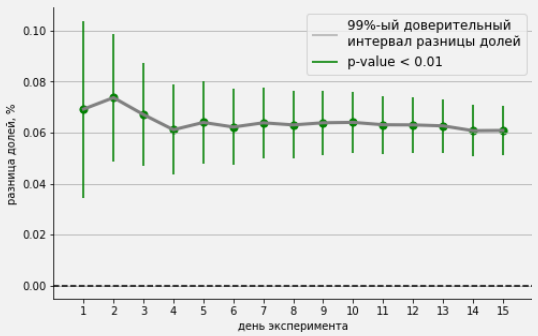
Проблема в другом. После предобработки данных в тестовой группе осталось всего 268 пользователей, дошедших до покупки. В одной выборке 268 средних чеков, в другой – 1693. Разница выборочных средних (331 у.е.) значительно ниже её стандартной ошибки (1315 у.е.). Опираться на тест при таких размерах выборок и таком незначительном различии – всё равно, что отвечать наугад.

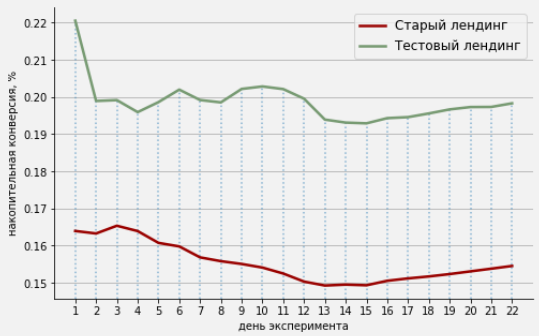
T-test с 20%-ой вероятностью ошибки второго рода и 5%-ым уровнем значимости будет обнаруживать эти различия тогда, когда в тестовой выборке наберётся 39 000 пользователей, дошедших до покупки, а в контрольной, соответственно, 351 000.

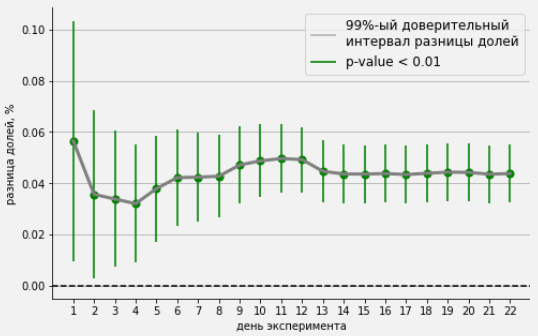
Гипотеза о различии в средних чеках не нашла подтверждения. Поэтому чтобы принять на основе данных решение, будет достаточно посчитать конверсию из посещения лендинга в первую покупку.







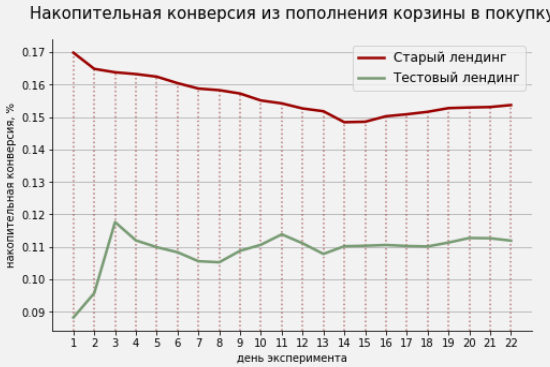


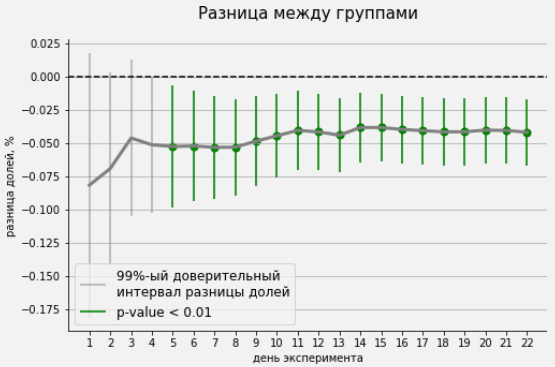


Принимаем все альтернативные гипотезы. В тестовой группе:

1. Bounce Rate значимо вырос на 6.09 пп.
2. Конверсия в пополнение корзины значимо выросла на 4.37 пп.
3. Конверсия из пополнения корзины в покупку значимо упала на 4.19 пп.

Доверительные интервалы, если они нужны, можно увидеть на графиках.





$$\text{Конверсия в первую покупку} = 1 \cdot (1 - \text{Bounce}) \cdot C_{\text{из активности в пополнение корзины}} \cdot C_{\text{из пополнения корзины в покупку}}$$

```
С текущей версии сайта до первой покупки доходят 10 из 1000 пользователей (0.00978)
С тестовой версии сайта до первой покупки доходят 8 из 1000 пользователей (0.00779)
```

Я советую оставить текущую версию сайта. Если команда верит в новый лендинг, то можно перезапустить A/B-эксперимент и провести его, исправив все недостатки. На случай, если тест окончится так же – пусть в отдельной таблице будет стратификация всех, кто участвует. Тогда можно будет изучить целевые метрики в разрезе подгрупп пользователей и, скорее всего, найти объяснение, почему они принимают не те значения, какие мы ждём.