Всё.Техника

Проведение и анализ A/B-теста

Дата анализа: 21.10.2022
Ишназарова Лера
tg: @leraish

Содержание

Описание проекта

Команда продакт-менеджеров маркетплейса “Всё.Техника” решила **выделить игровые ноутбуки в отдельную категорию товаров** (до этого они были в одной категории с ПК и всеми ноутбуками). Предполагается, что пользователи не могут найти игровые ноутбуки среди всей остальной компьютерной техники.

Прежде, чем внедрить изменение, решено провести A/B-тестирование. Перед запуском эксперимента определены:

- **целевые метрики:**

  конверсия в покупку - в тестовой группе ожидается прирост на 100%
  средний чек - ожидается, что в тестовой группе останется на том же уровне (метрика не изменится)

- **аудитория:**

  только новые пользователи

Остальные параметры теста необходимо рассчитать, опираясь на исторические данные. После запуска эксперимента необходимо оценить корректность проведения теста, а после его остановки проанализировать результаты.

План исследования

План состоит из трех хронологических блоков: подготовка к эксперименту - расчет оставшихся параметров теста и составление ТЗ, запуск эксперимента - оценка корректности проведение теста, остановка эксперимента - анализ результатов теста

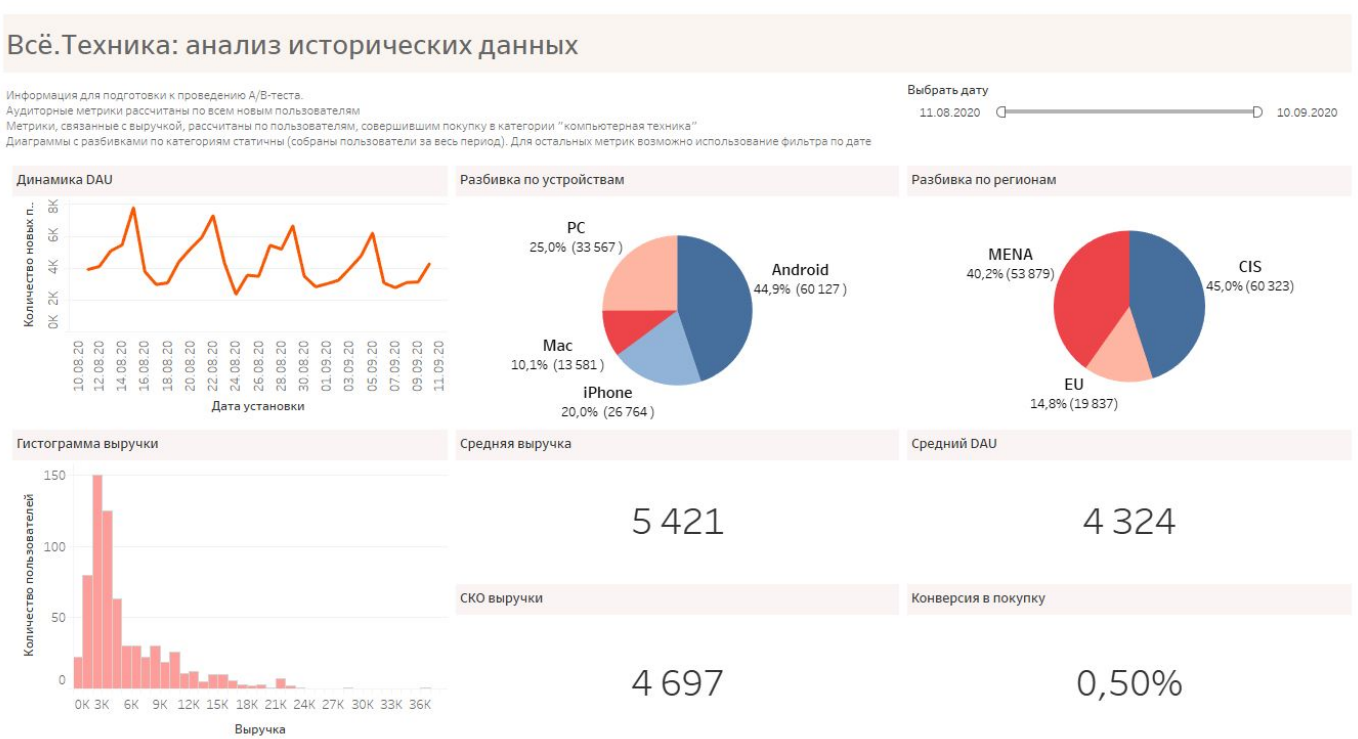
I. **Расчет параметров теста:**
- Получение и анализ исторических данных (DAU новых пользователь за предыдыщие 4 недели, выручка и ее параметры по пользователям, совершившим покупку в категории “компьютерная техника”)
- Разработка дашборда с интересующими метриками
- Определение оставшихся входных параметров теста: минимальный размер групп, длительность теста, дата его запуски и остановки
- Составление ТЗ для проведения A/B-теста, формулирование гипотез, выбор стат.методов их проверки

II. **Оценка корректности проведения теста**
- Получение данных по проводимому тесту (баланс групп)
- Разработка дашборда с данными об участниках теста
- Проверка корректности проведения теста с помощью стат.тестов
- Принятие решение об остановке или продолжении эксперимента

III. **Анализ результатов эксперимента**
- Получение данных о результатах теста (баланс групп, интересующие метрики)
- Разработка дашборда с целевыми метриками в разбивке по группам
- Анализ различий между группами, оценка этих различий с помощью стат.тестов
- Выводы и принятие решения о внедрении изменений для всей аудитории

Расчет параметров теста

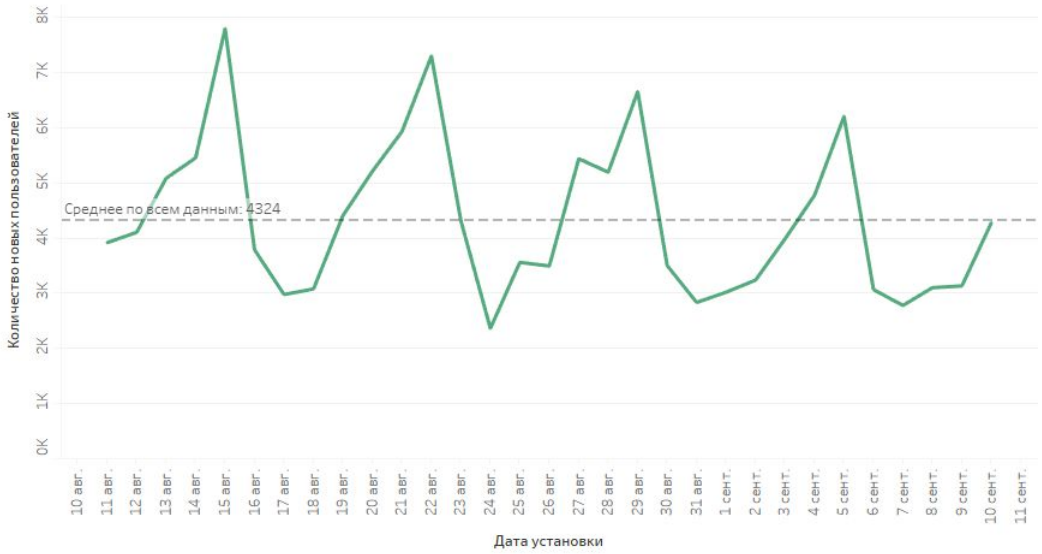
Исторические данные: дашборд



Открыть в Tableau

DAU

Динамика DAU новичков



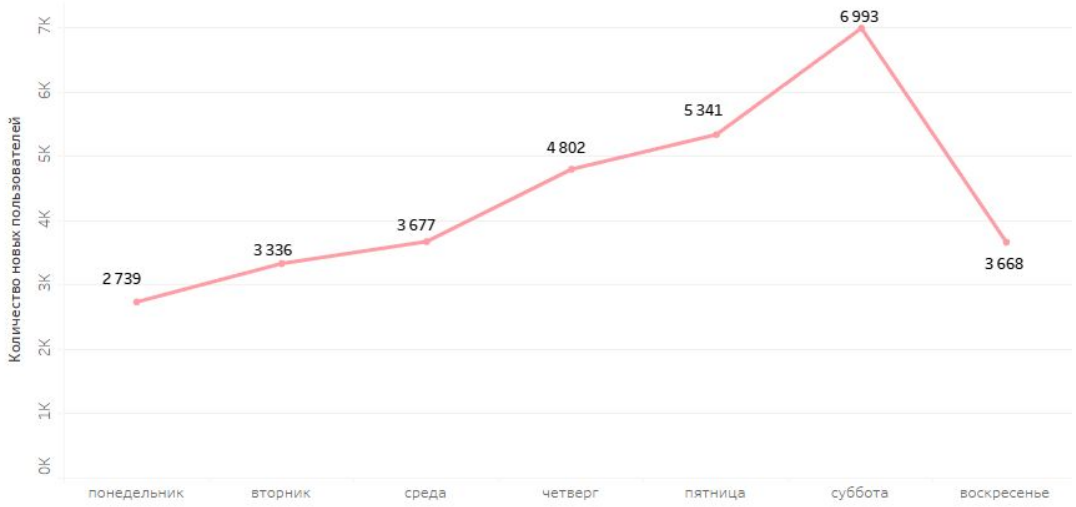
Исследованы данные по новым пользователям за предыдущие 4 недели (11.08.2020 - 10.09.2020).

Среднее значение DAU: 4 324.

Наблюдаем определенный паттерн - пиковые значения приходятся на субботу.

Недельная сезонность DAU

Средний DAU по дням недели

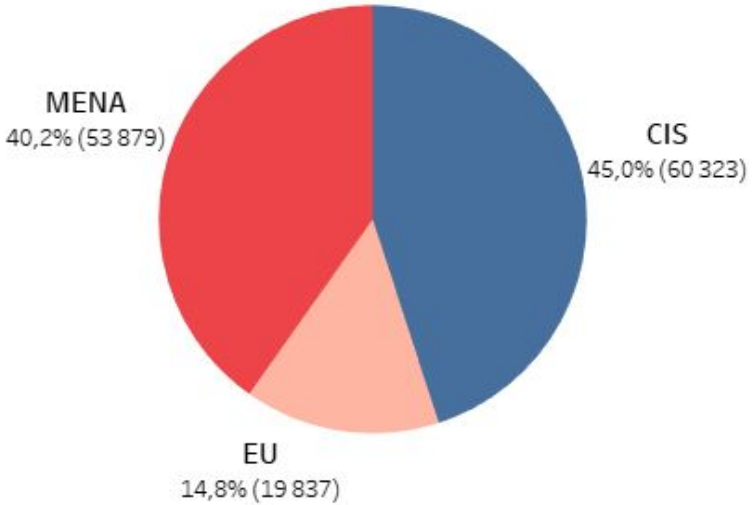


При оценке среднего DAU по дням недели ярко выражается **недельная сезонность:** количество активных пользователей постепенно увеличивается каждый будний день, достигая пика к **субботе.** В воскресенье аудитория сокращается почти вдвое.

Важно учитывать выявленную особенность при определении сроков эксперимента.

Распределение по регионам

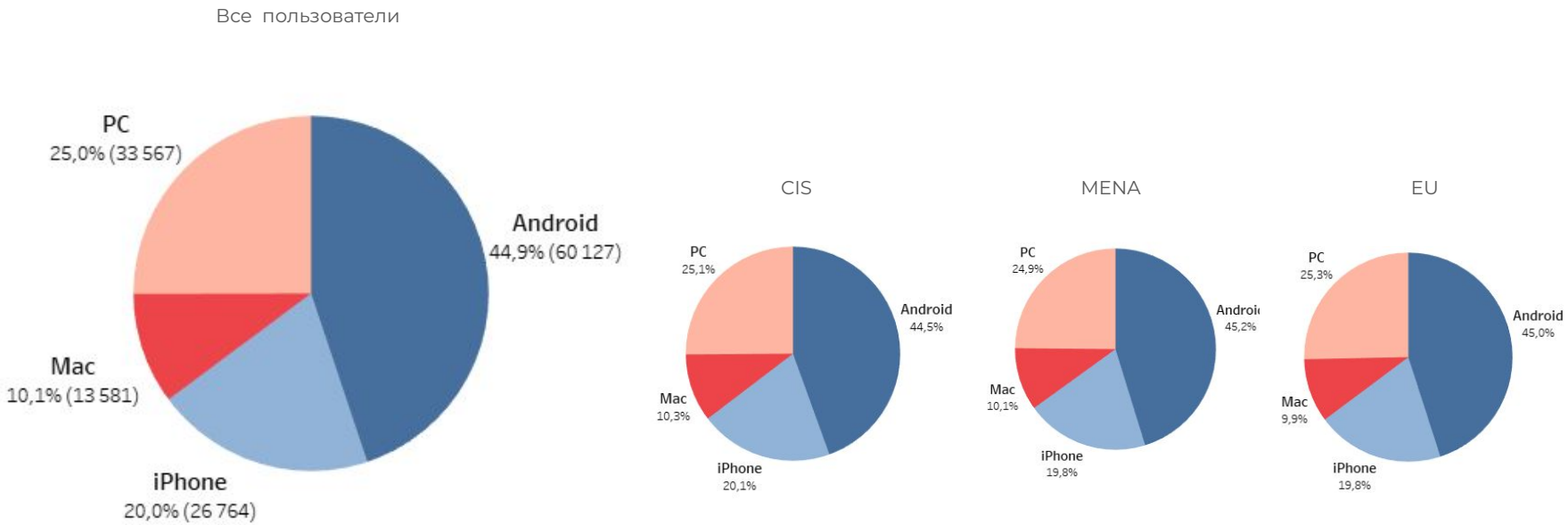
Доли пользователей по регионам



Территориально пользователи распределены **неравномерно**.

Количество клиентов из регионов MENA и CIS относительно сопоставимо (45% и 40.2%), пользователей из EU всего 14.8%

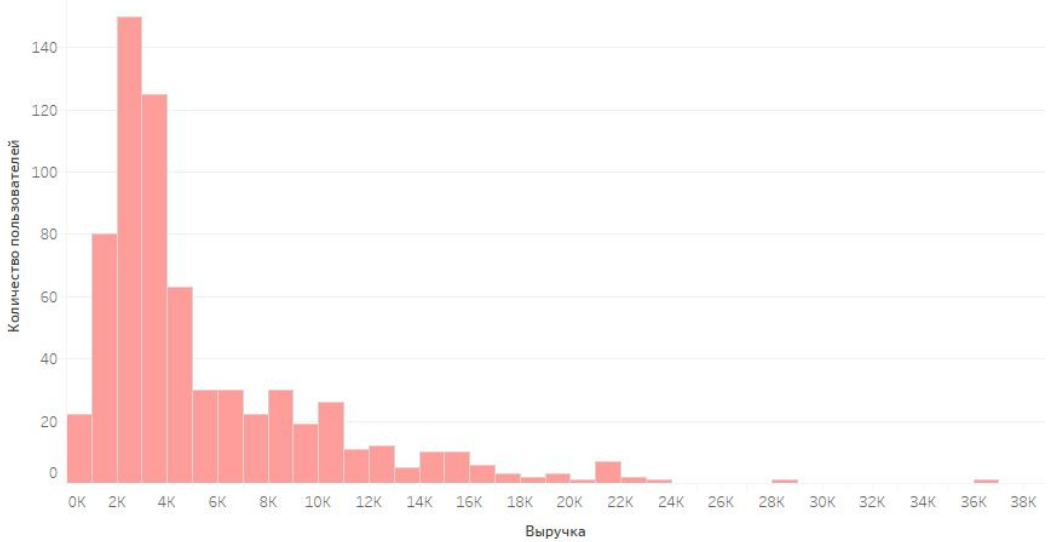
Распределение по устройствам



Большинство пользуется Android. **Доли устройств в разных регионах практически не отличаются.**

Выручка и конверсия

Гистограмма выручки



Средняя выручка

5 421

СКО выручки

4 697

Конверсия в покупку

0.5%

На гистограмме выручки наблюдаем характерное для метрики **гамма-распределение:** длинный “хвост” вправо, немного экстремально больших значений.

Наличие таких больших значений оказывает влияние на среднее (5 421 при медиане 3 649) и СКО, но исключать эти данные из анализа не стоит - такое распределение для монетарных метрик является нормой. **Ожидаем наличие схожих выбросов при проведении теста, которое необходимо учесть при оценке изменений в среднем чеке.**

Наблюдения по историческим данным

DAU

- средний DAU - 4 324 пользователя
- прослеживается недельная сезонность (аудитория нарастает к субботе. к воскресению падает)

Распределение по регионам и устройствам

- MENA: 40.2% пользователей, CIS 45%, EU 14.8%
- Большинство (45%) используют Android. Распределение по устройствам одинаково для всех регионов

Выручка и конверсия

- средний чек 5421, СКО 4 697. Встречаются аномально высокие чеки, что нормально для метрики
- Конверсия в покупку 0.5%

Определение параметров теста

Минимальный размер выборки

- Для наблюдения изменений в **конверсии: 4 673**
  - *Исторический уровень конверсии: 0.5%*
  - *Ожидаемый эффект: +100%*
  - *Среднее количество участников в неделю: 30 268 (средний DAU*7)*
  - *Двусторонняя гипотеза, уровень значимости 0.5%, мощность теста 80%*
  - *Калькулятор Вычисления*
- Для наблюдения изменений в **среднем чеке: 4 716**
  - *Историческое значение: 5 421*
  - *Ожидаемый эффект: метрика не изменится, учитываем изменения на 5% - 5 692*
  - *СКО: 4 697*
  - *Двусторонняя гипотеза, уровень значимости 0.5%, мощность теста 80%*
  - *Калькулятор Вычисления*

Выбираем бОльший размер: **4 673 участников в каждой группе**

Длительность теста

(Размер выборки / Средний DAU) * количество групп = (4 716 / 4 324) * 2 = 2,2
Учитывая обнаруженную недельную сезонность, округляем до **7 дней**

Дата запуска

Учитывая праздники и маркетинговые мероприятия: **14 октября 2020**

Расчет параметров теста

Техническое задание на проведение теста

Название теста: gaming_laptops_test

Назначение теста: выделение игровых ноутбуков в отдельную категорию товаров для улучшения поиска

Аудитория: только новые пользователи, без привязки к конкретному региону и/или устройству

Тестовые группы и их размер: current_group (контрольная) и new_group (тестовая). Минимум 4 676 участников в каждой из групп

Сроки проведения теста, длительность: 14.10.2020 - 20.10.2020, 7 дней

Тестируемые метрики:

- **Конверсия в покупку.** Ожидаемый эффект: +100% в тестовой группе
- **Средний чек.** Ожидаемый эффект: в тестовой группе метрика не изменится относительно контрольной

Техническое задание на проведение теста

Проверяемые гипотезы:

Конверсия:

H0: Между конверсиями тестовой и контрольной групп НЕТ различий
H1: Между конверсиями тестовой и контрольной групп ЕСТЬ различия

Проверка двухвыборочным z-тестом для пропорций. Ожидается подтверждение H1.

Средний чек:

**Для проверки t-тестом при отсутствии выбросов*
H0: Между средним чеком тестовой и контрольной групп НЕТ различий
H1: Между средним чеком тестовой и контрольной групп ЕСТЬ различия

**Для проверки тестом Манна-Уитни, при наличии выбросов (+визуальная оценка)*
H0: Данные в выборках контрольной и тестовой групп получены из ОДНОЙ генеральной совокупности
H1: Данные в выборках контрольной и тестовой групп получены из РАЗНЫХ генеральных совокупностей

В обоих случаях ожидается подтверждение H0

Для всех проводимых тестов установлен уровень значимости 0.05 и мощность теста 80%

Оценка корректности проведения теста

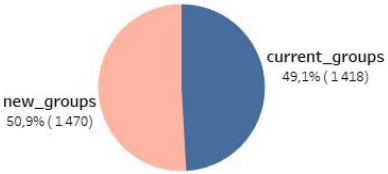
Проведение теста: дашборд

Всё Техника: проведение A/B-теста gaming_laptop_test

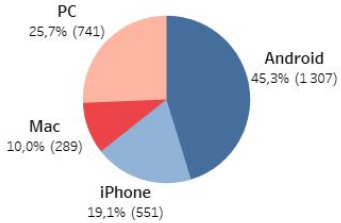
Динамика набора участников тестирования.
Тест запущен 14.10.2020, данные обновляются ежедневно в 00:00 МСК

Выбрать дату: 14.10.2020 — 14.10.2020
Выбрать группу: (All)
Выбрать регион: (All)
Выбрать устройство: (All)

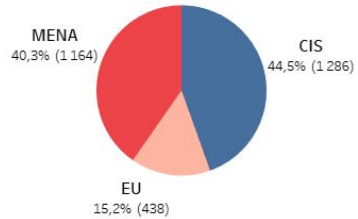
Баланс групп



Разбивка по устройствам



Разбивка по регионам



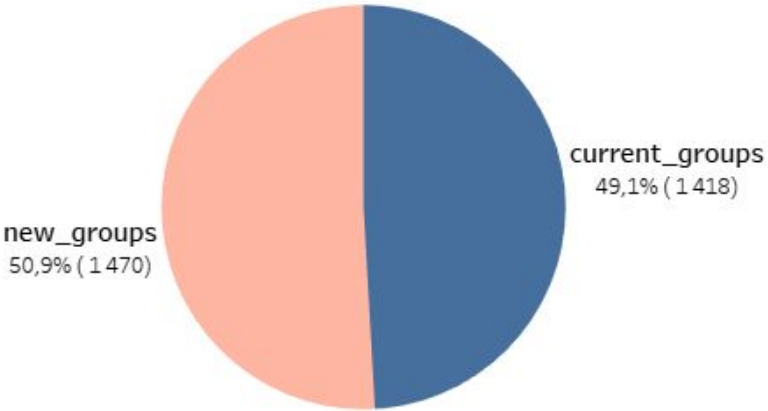
Число участников теста по регионам и устройствам

Устройство	CIS		MENA		EU	
	current_groups	new_groups	current_groups	new_groups	current_groups	new_groups
Android	47,9% 309	46,3% 297	44,0% 245	45,0% 273	44,9% 97	38,7% 86
PC	24,8% 160	25,4% 163	26,0% 145	23,4% 142	32,4% 70	27,5% 61
iPhone	17,1% 110	18,7% 120	20,8% 116	20,4% 124	14,8% 32	22,1% 49
Mac	10,2% 66	9,5% 61	9,2% 51	11,2% 68	7,9% 17	11,7% 26

Открыть в Tableau

Баланс групп

Количество участников теста по группам



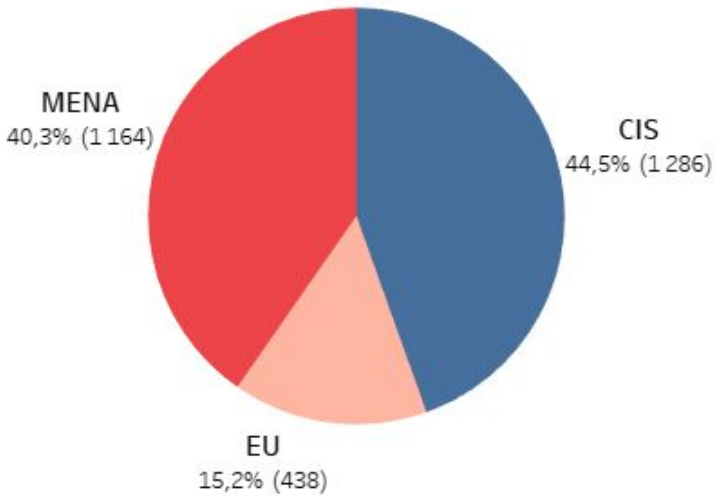
Тест запущен 14.10.2020, за первый день удалось набрать **2 888 участников**. Учитывая исторические данные, для четверга ожидалось бОльшее количество, однако при удержании даже такого темпа набора участников к дате окончания теста соберется необходимое количество.

Видим небольшой перевес в сторону тестовой группы, **но разница статистически не значима**: проведен z-тест, подтверждающий равномерное распределение по группам (при уровне значимости 0.05 p-value 0.49)*

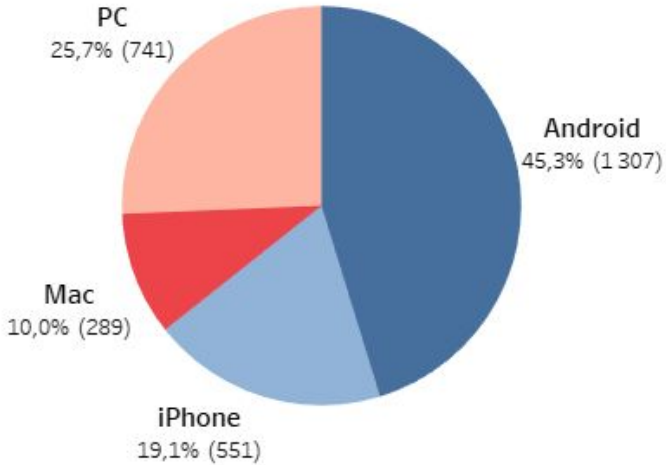
**Калькулятор*
Вычисления

Разбивка по регионам и устройствам

Количество участников теста по регионам



Количество участников теста по типам устройства



Разбивка пользователей по регионам и устройствам **соответствует историческим данным,** пропорции практически не отличаются.

Разбивка по регионам и устройствам

Устройство	CIS		MENA		EU	
	current_groups	new_groups	current_groups	new_groups	current_groups	new_groups
Android	47,9% 309	46,3% 297	44,0% 245	45,0% 273	44,9% 97	38,7% 86
PC	24,8% 160	25,4% 163	26,0% 145	23,4% 142	32,4% 70	27,5% 61
iPhone	17,1% 110	18,7% 120	20,8% 116	20,4% 124	14,8% 32	22,1% 49
Mac	10,2% 66	9,5% 61	9,2% 51	11,2% 68	7,9% 17	11,7% 26

В каждом регионе набираем пользователей с разными устройствами, **доли между регионами схожи и не противоречат историческим данным.**

По группам теста внутри региона есть несколько выделяющихся отличий (например, перевес в тестовой группе iPhone и Mac в EU), что в целом нормально для малочисленной выборки.

Экстремальных различий нет.

Наблюдения и выводы по проведению теста

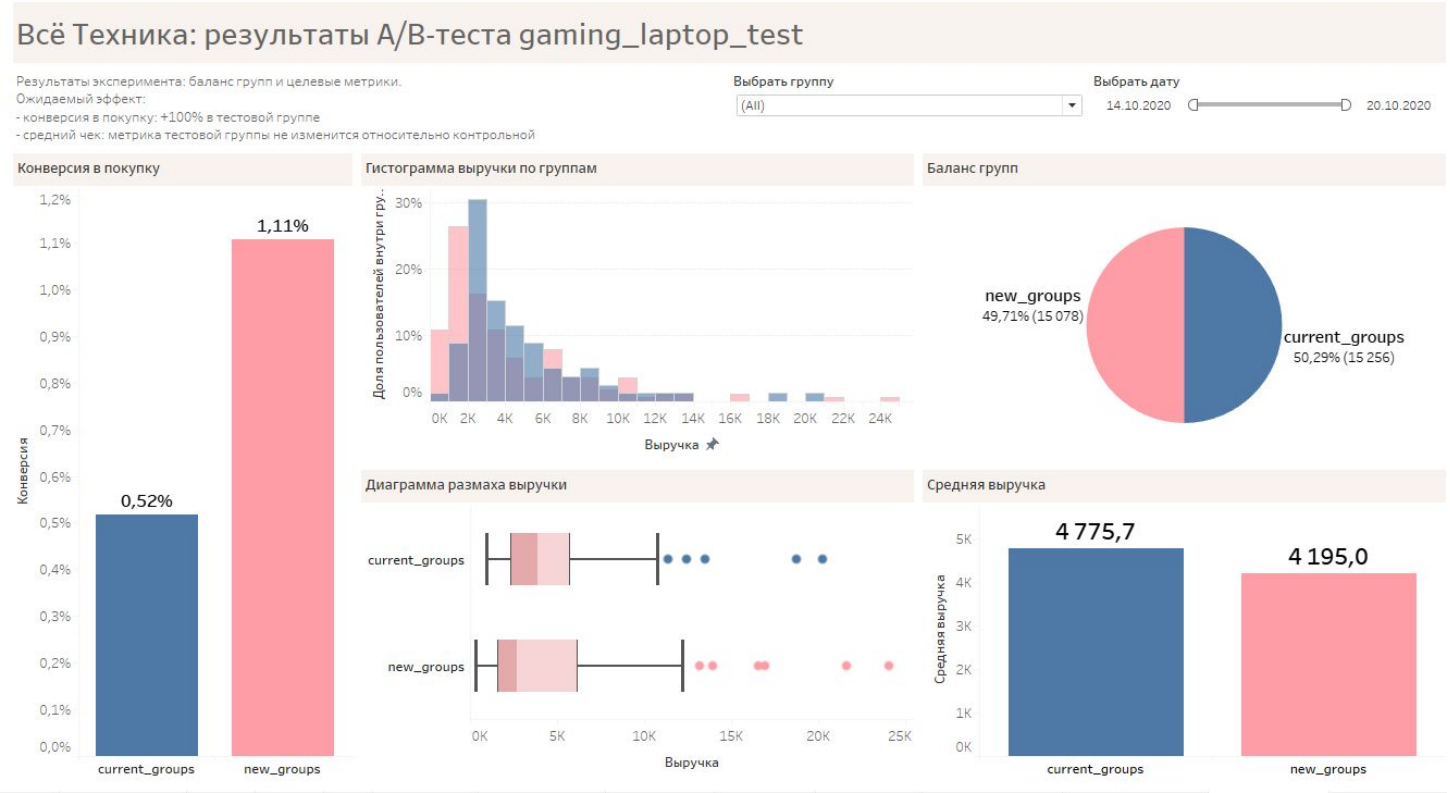
- За первый день эксперимента удалось собрать 2 888 участников, темп набора нормальный
- Участники теста набираются равномерно по группам - 50.9% в тестовой, 49.1% в контрольной (разница статистически не значима)
- Доли отобранных участников в разбивке по регионам и типам устройства не противоречат историческим данным
- Внутри каждого региона набираем пользователей с разными типами устройств в соответствии с историческими данными.

Данные собираются корректно, причин для аварийной остановки теста нет.

Тест продолжается до момента достижения установленного размера выборки и заданной длительности.

Анализ результатов эксперимента

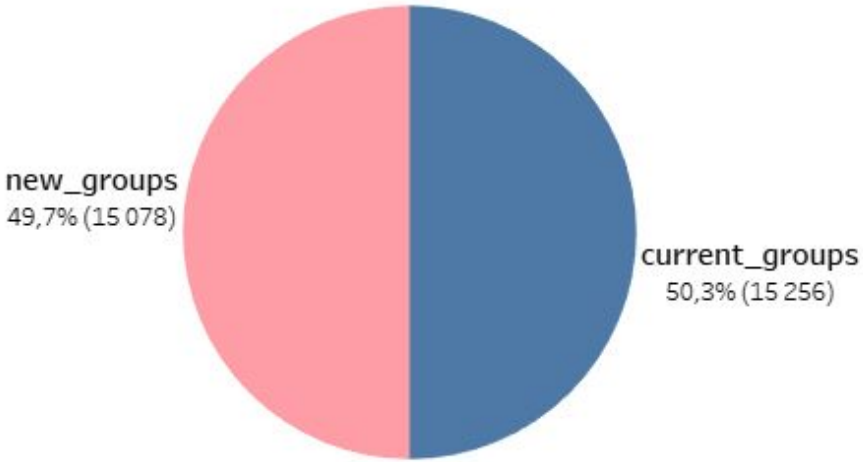
Результаты теста: дашборд



Открыть в Tableau

Остановка теста

Количество участников по группам теста



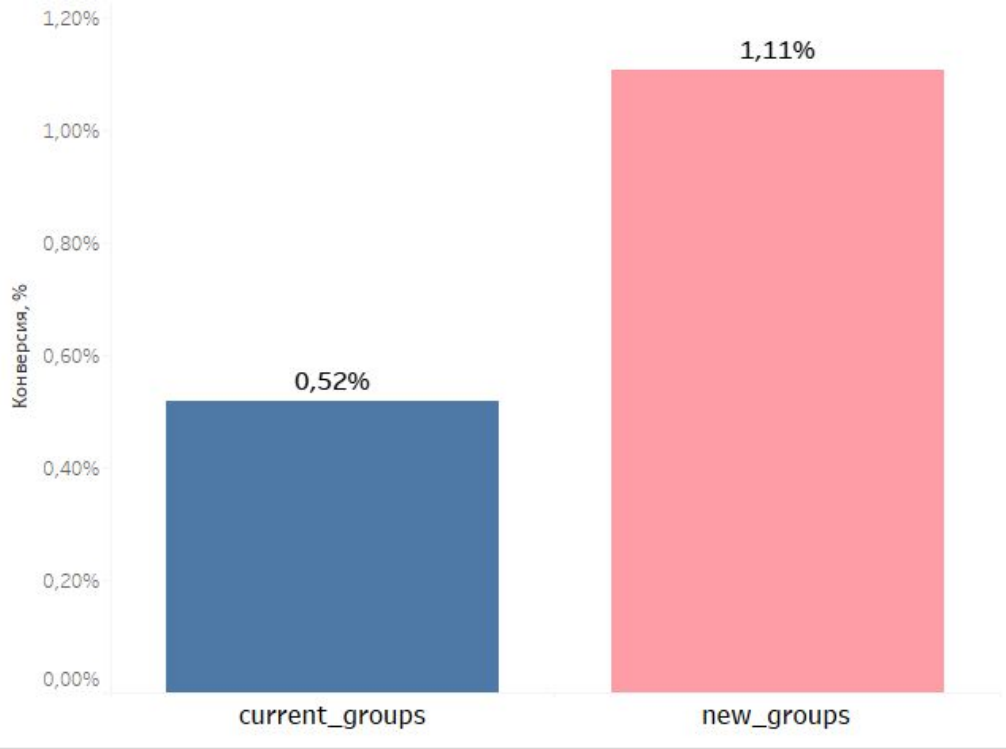
- достигнута заданная продолжительность эксперимента (7 дней, тест проводился 14.10.20 - 20.10.20)
- набран установленный минимальный размер выборки
- каждый пользователь попал только в одну группу теста
- участники равномерно распределены по группам (проверка пропорции z-тестом, p-value(0.46) > 0.05)*

Основные критерии соблюдены, тест успешно остановлен

**Калькулятор Вычисления*

Конверсия в покупку. Z-тест

Конверсия в покупку по группам A/B-теста



Конверсия тестовой группы превышает конверсию контрольной. Разница статистически значима: проведен двухвыборочный z-тест* для пропорций, p-value(0) меньше уровня значимости 0.05, т.е. эффект не случайный.

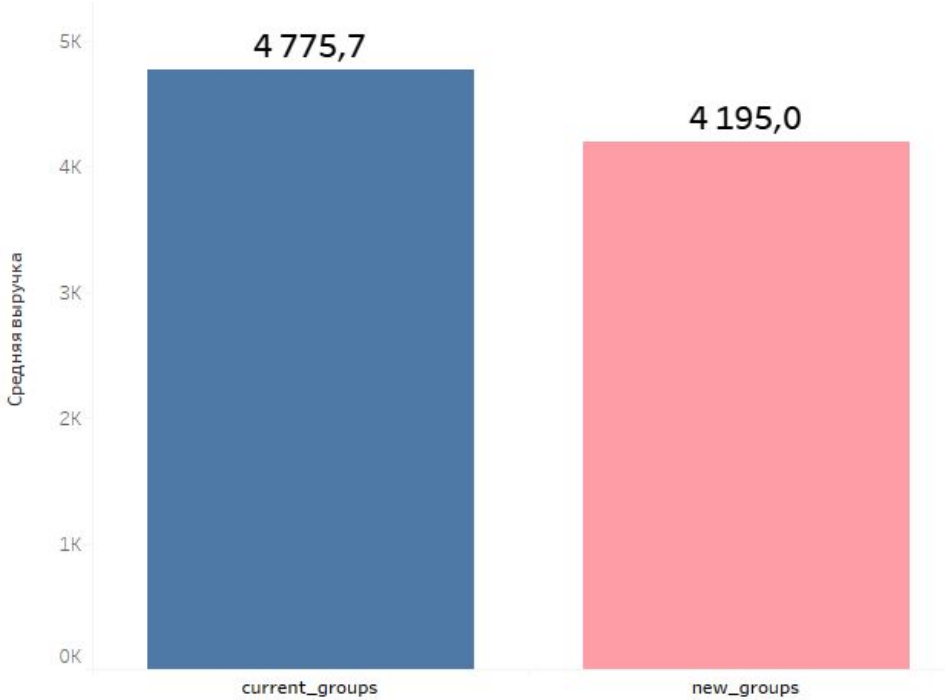
Выдвинутая перед запуском эксперимента H0 не подтверждается, между группами ЕСТЬ стат. значимые различия.

Ожидаемый эффект (+100% в тестовой группе) достигнут.

*Калькулятор
Вычисления

Средний чек. Т-тест

Средний чек по группам A/B-теста



Арифметически **средний чек тестовой группы ниже на 580.7** относительно контрольной группы, что противоречит нашим ожиданиям.

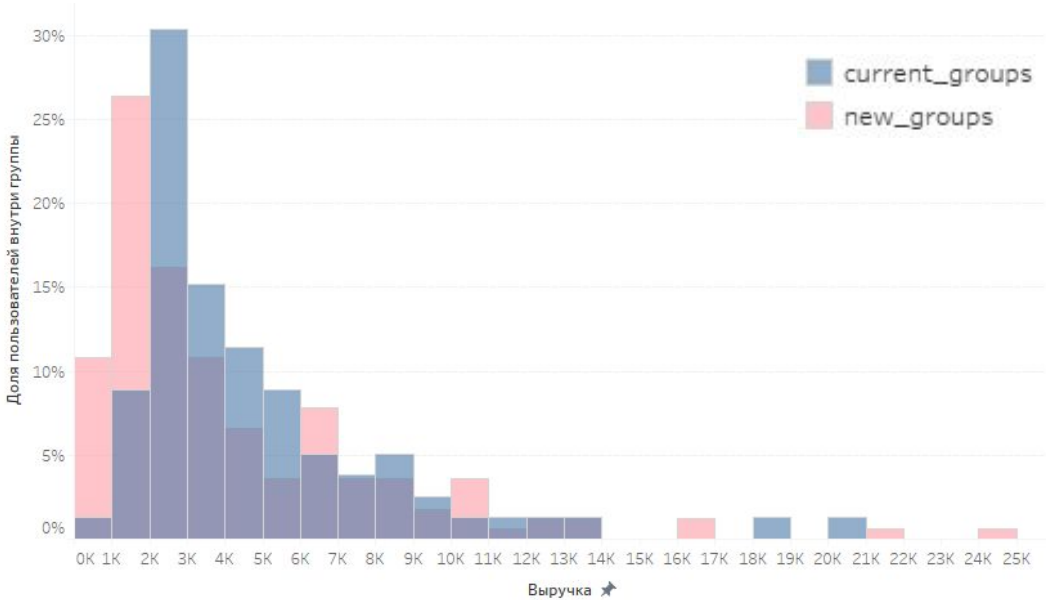
Проверка результатов **t-тестом*** не отвергает выдвинутую H0 - p-value(0.27)>уровня значимости, следовательно между средними чеками нет стат. значимой разницы.

Ранее при анализе исторических данных выявили, что метрике свойственны выбросы, высокий уровень СКО и наличие экстремально высоких чеков. Необходимо сравнить результаты групп между собой, проанализировать существующие выбросы и при необходимости **проверить группы тестом Манна-Уитни.**

**Калькулятор*
Вычисления

Анализ выручки

Гистограмма выручки по группам A/B-теста



“Хвост” гистограммы участников тестовой группы длиннее, чем у гистограммы контрольной группы, самые дорогие покупки совершили пользователи из new_groups.

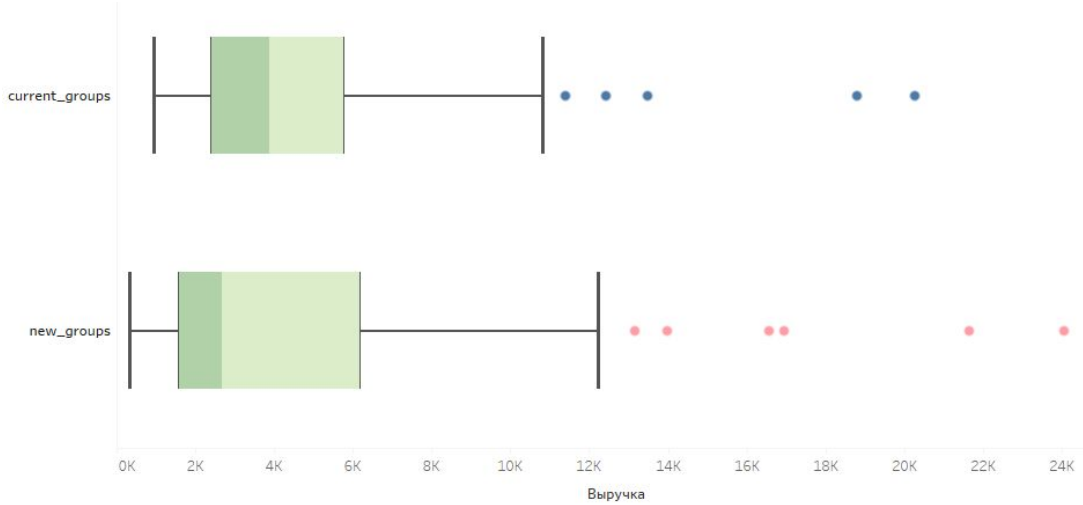
При этом **средний чек тестовой группы чаще всего располагается в пределах 1 000-2 000,** в то время как для контрольной группы самый частый диапазон 2 000-3 000.

Ярко выражена доля **крайне низких чеков** тестовой группы (0-1 000), у контрольной группы эта доля значительно ниже.

Высокий уровень конверсии связан большим количеством “дешевых” покупок.

Выручка. Размах данных

Диаграмма размаха выручки по группам A/B-теста



Медианное значение выручки контрольной группы выше, чем в тестовой (3 864 против 2 682). Также у контрольной группы “левый ус” и нижняя граница “ящика” расположены правее относительно тех же значений тестовой группы **(у тестовой группы численно больше низких значений**)

При этом правый “ус” тестовой группы расположен дальше, у группы больше выбросов, больше “дорогих” покупок.

Итого, у тестовой группы численно больше как низких, так и высоких чеков, но судя по среднему, **количество дорогих покупок не компенсирует количество дешевых.**

Тест Манна-Уитни. Вывод по среднему чеку

Т.к. обе выборки подвержены выбросам, которые не следует исключать из анализа, проведен тест Манна-Уитни*, позволяющий проверить получены ли выборки из одной генеральной совокупности.

Вычисленное P-value (0.01016) меньше уровня стат.значимости, следовательно, выборки получены из разных генеральных совокупностей. **Их параметры отличаются.**

Дополнительно проверен средний чек одной транзакции** - так же нельзя сделать вывод о том, что выборки получены из одной генеральной совокупности (p-value 0.00001)

Резюмируя результаты теста и полученные средние значения: **средний чек тестовой группы изменился относительно контрольной группы в худшую сторону. Ожидаемый эффект не достигнут.**

*****_Калькулятор Вычисления_
** Выручка с одного клиента поделена на количество сделанных им транзакций Калькулятор Вычисления

Выводы и рекомендации по результатам эксперимента

Тест запущен 14.10.2022 и успешно остановлен 20.10.2022. Участвовало **30 334 пользователей**, равномерно распределенных по двум группам. Соблюдены все критерии успешной остановки теста.

Результаты эксперимента проанализированы и разница между группами проверена статистическими тестами:

➔ **Конверсия в покупку:**

Ожидаемый эффект (+100% в тестовой группе) достигнут. Конверсия тестовой группы **1.11%**, конверсия контрольной группы 0.52%, историческая конверсия: 0.5%.

Результат эксперимента проверен и подтвержден z-тестом для пропорций, эффект не случайный, введение отдельной категории игровых ноутбуков **положительно сказалось на конверсию в покупку** во всей категории “Компьютерная техника”

➔ **Средний чек:**

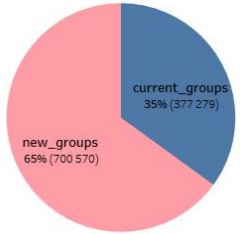
Ожидаемый эффект (метрика не изменится) не достигнут. Средний чек тестовой группы **4 195,** что на 580.7 меньше, чем в контрольной.

Т.к. метрика подвержена выбросам, содержит в себе экстремально высокие значения, группы проверены тестом Манна-Уитни - по его результатам, выборки получены из разных генеральных совокупностей, т.е. статистические различия есть.

Медианное значение среднего чека тестовой группы также ниже - 2 682 против 3 864 в контрольной. Большинство значений среднего чека в тестовой группе расположены в диапазоне 1 000 - 2 000, велика доля чеков до 1 000 - как и внутри тестовой группы, так и относительно контрольной.

Выводы и рекомендации по результатам эксперимента

Стоит ли внедрять разделение игровых ноутбуков к внедрению для всей аудитории?



Несмотря на то, что средний чек существенно понизился, необходимо ответить на вопросы:

➔ **компенсирует ли повышение одной метрики (конверсии) падение другой (среднего чека)?**
чисто математически - да, компенсирует. 65% выручки мы получили именно от тестовой группы

➔ **потеряли ли мы в продажах игровых моделей?**
С одной стороны, внедрив подкатегорию, мы облегчили поиск пользователям, нацеленным купить именно игровой ноутбук; с другой, могли получить следующие пути пользователя:

- клиент не замечает новую подкатегорию
- клиент не ориентируется в отличиях игровой/обычный ноутбук, **намеренно игнорирует** новую подкатегорию. Возможно, само понятие “игровой” отталкивает от покупки, в то время как в общей категории такие модели рассматривались бы наравне со всеми

В обоих гипотетических случаях клиент ограничивается только общей категорией и выбирает из более дешевых товаров - возможно, с этим связано “засилье” низких чеков в тестовой группе. **Необходимо проанализировать:**

- **Путь пользователя тестовой группы:** как часто заходили на новую страницу “игровые ноутбуки”? Переходили ли на эту страницу после страницы со всей компьютерной техникой?
- **Сравнение конверсии в покупку игрового ноутбука между группами теста:** действительно ли мы теряем потенциальных покупателей игровых моделей?
- **Сравнение конверсии тестовой группы в покупку игрового ноутбука с историческими данными:** упали ли значения? Потеряли ли мы в доходе именно в этой подкатегории?

Выводы и рекомендации по результатам эксперимента

➔ **какой положительный эффект мы получили? Как мы можем использовать эти наблюдения в дальнейшем? Внедрять ли разделение на всю аудиторию?**

Как минимум, конверсия и выручка тестовой группы существенно выше, чем у контрольной. Финансовый эффект положительный, но неизвестно как долго мы сможем его удерживать - всегда ли высокая конверсия сможет компенсировать низкий чек?

Если мы выявим, что существенно потеряли в продажах игровых ноутбуков, то внедрять разделение не стоит. Но из результатов эксперимента можем найти иные точки роста:

- **Выделилась ли положительно какая-то другая подкатегория в тестовой группе?** Можем ли мы сконцентрироваться на ней? Можно сравнить конверсии и средние чеки в подкатегориях между группами теста и с историческими данными
- **Есть ли у пользователей запрос на разделение товаров по ценовым категориям?** Если нам финансово выгоднее продавать более дешевые модели, возможно стоит выделить именно их в отдельную группу?
- **Корректно ли у нас работают фильтры и сортировки?** Вероятно, вместо внедрения отдельной категории разумнее внедрить (модернизировать текущий) фильтр, сделав его более очевидным и понятным.