1. Дизайн эксперимента

• Цель эксперимента

Выявить при какой модели расчета ставки («Максимальная Ставка» или «Средняя Ставка») достигается наиболее эффективное привлечение клиентов для покупки товаров у компании «eStore».

• Нулевая и альтернативная гипотезы

H0: Средние по тестовой и контрольной группе равны (изменения модели расчета ставки ни к чему не приведут)
H1: Средние по тестовой и контрольной группе неравны (изменения модели расчета ставки приведут к росту или падению покупок)

• Определение изменения и целевые метрики

- Spend [USD] (затраты на рекламу)
- # of Website Clicks (количество кликов)
- # of Purchase (количество приобретений пользователями товаров)
- actions (количество действий = количество покупок + количество поисков на сайте + количество просмотров деталей товара + количество добавлений в корзину)
- CR (конверсия) = actions / количество кликов
- CPO (стоимость за заказ) = Расходы / Количество покупок

• Экспериментальная и контрольная группы

Для экспериментальной группы применяется модель расчета «Средняя Ставка», а для контрольной группы оставляем «старую» модель расчета - «Максимальная Ставка».

• Уровень статистической значимости

Альфа = 0,05, Мощность = 0,8

• Параметры эксперимента

- Размер выборки. Уникальных пользователей, посмотревших рекламу, в тестовой группе - 1509609, в контрольной группе - 2576503, всего - 4086112.

- Факторы воздействия. Сезонный фактор, окно принятия решений

- Длительность эксперимента. 30 дней

2. Проведение и анализ эксперимента

Шаг 1. Проверка разделения наблюдений на экспериментальную и контрольную группы на основе данных, полученных после проведения эксперимента.

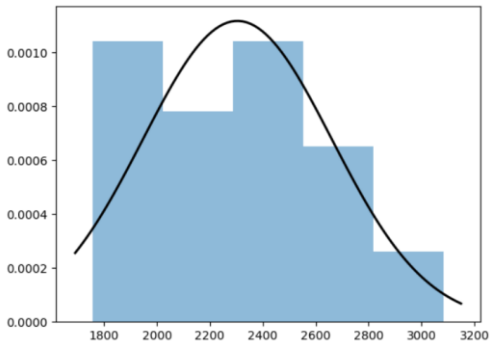
Сформированные данные по тестовой и контрольной группах были проверены на пропущенные значения. В результате проверки обнаружились пропущенные данные за 5 августа 2019 года в контрольной группе. Была рассчитана доля пропущенных значений, которая оказалась несущественной в общем объеме данных по каждому из исследуемых показателей (примерно 3%), поэтому было принято решение удалить данные за 5 августа 2019 года в тестовой и контрольной группах для того, чтобы эксперимент был более чистым.

Шаг 2. Подсчет метрик на основе данных, полученных после проведения эксперимента. Проверка нормальности распределения.

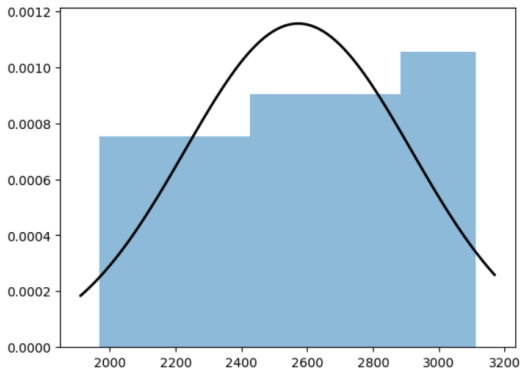
Для достижения цели эксперимента нужно выбрать такие метрики, которые отражали бы заинтересованность покупателей рекламным предложением и количество совершенных ими целевых действий. Кроме того, бизнесу важно следить и за затратами на рекламу. Так, в качестве метрик, были выбраны:

- Spend [USD] (затраты на рекламу)
- # of Website Clicks (количество кликов)
- # of Purchase (количество приобретений пользователями товаров)
- actions (количество действий = количество покупок + количество поисков на сайте + количество просмотров деталей товара + количество добавлений в корзину)
- CR (конверсия) = actions / количество кликов
- CPO (стоимость за заказ) = Расходы / Количество покупок

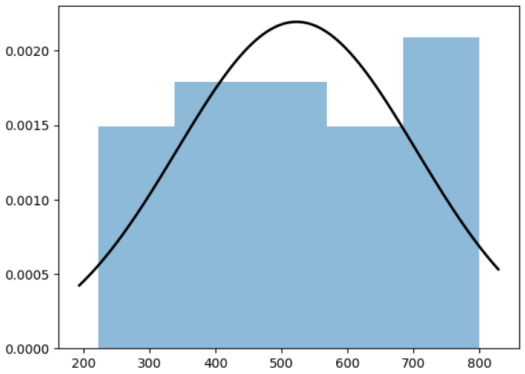
Для проверки нормальности распределения проведен тест Шапиро-Уилка и построены графики теоретических и эмпирических частот.



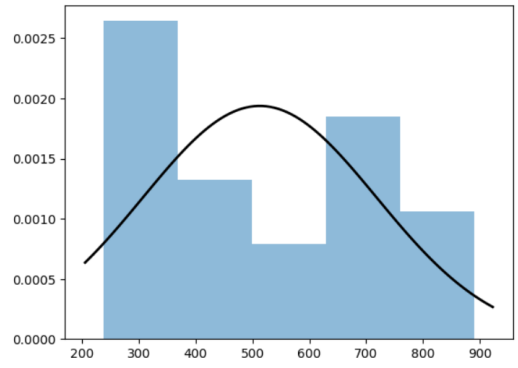
Затраты на рекламу, группа А, распределение нормальное



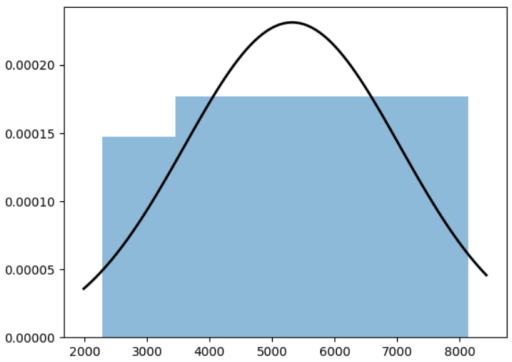
Затраты на рекламу, группа Б, распределение нормальное



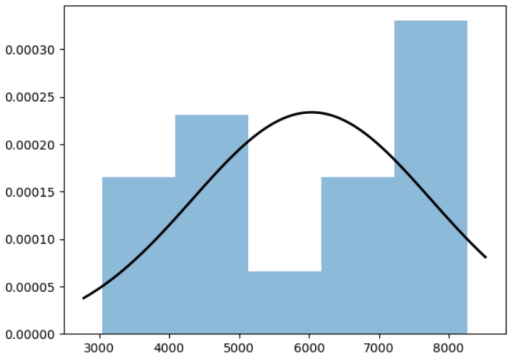
Количество покупок, группа А, распределение нормальное



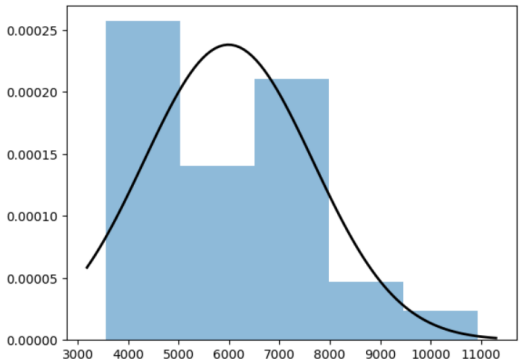
Количество покупок, группа Б, распределение не нормальное



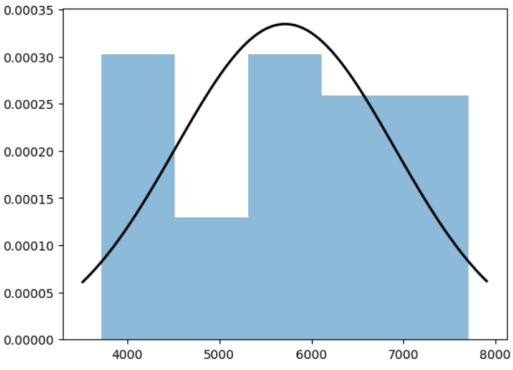
Количество кликов, группа А, распределение нормальное



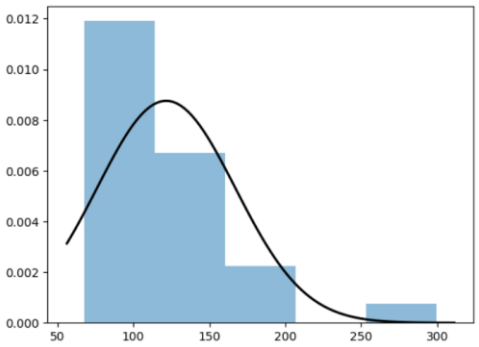
Количество кликов, группа Б, распределение не нормальное



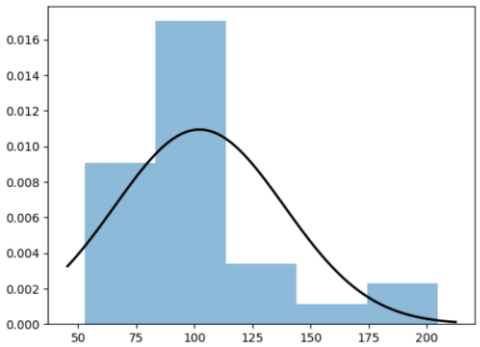
Количество действий, группа А, распределение нормальное



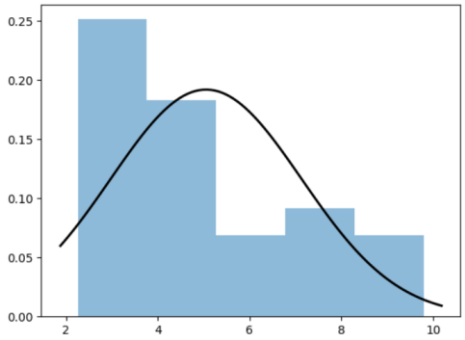
Количество действий, группа Б, распределение нормальное



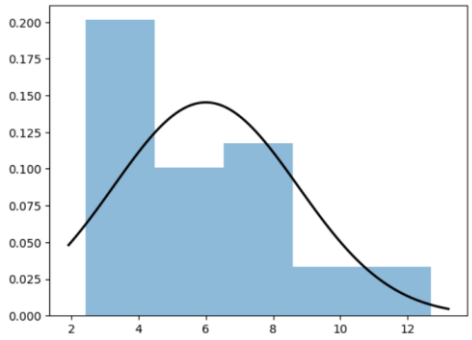
Конверсия, группа А, распределение не нормальное



Конверсия, группа Б, распределение не нормальное



СРО, группа А, распределение не нормальное



СРО, группа Б, распределение не нормальное

Шаг 3. Формулирование гипотезы. Подбор статистического критерия для тестирования гипотезы

Для метрик, которые имеют нормальное распределение (затраты на рекламу, количество действий), будем использовать двувыборочный критерий Стьюдента. Таким образом,

H0: средние значения двух выборок одинаковы
H1: не одинаковы

Так, по p-значению в тесте Стьюдента можно сказать, что для затрат на рекламу мы отклоняем нулевую гипотезу, т.к. p-значение меньше 0,05, а для метрики «количество действий» мы принимаем нулевую гипотезу о равенстве средних.

Другие метрики будем проверять критерием Манна-Уитни, поскольку в нём не заложены представления о базовой форме распределения. Таким образом,

H0: средние значения двух выборок одинаковы
H1: не одинаковы

Так, в тесте Манна-Уитни мы для каждой метрики принимаем гипотезу о равенстве средних по двум выборкам, поскольку p-значение больше 0,05.

Шаг 4. Определение параметров для A/B групп. Проведение A/A теста для оценки параметров. Проверка репрезентативности и однородности выборок

Поскольку в рамках кейса нам даны данные по одной тестовой и по одной контрольной группе, для которых использовались разные методы расчета ставки в течение 30 дней, проведение A/A теста не представляется возможным. Вместо этого используем процесс ресэмплинга для создания сравнительных групп, которые должны быть однородными и подобными первоначальным группам.

Мы провели бутстрап ресэмплинг отдельно сформировав новые выборки и оценив средние значения по этим выборкам для группы A и для группы B. В итоге получилось, что процент p-значений, которые меньше 0,05 для каждой метрики в группе A и в группе B находится в районе уровня значимости и в некоторых случаях пересекает его. Это говорит о том, что контрольные и тестовые группы могут быть неоднородны или требуется проводить эксперимент дольше, чтобы собрать больше данных. Также хотелось бы провести исследование на данных не за определенное количество дней, а на датасете, в котором была бы информация (целевые действия, время и дата посещения, личные характеристики (пол, возраст, география) и др.) по каждому пользователю. Так, данных было бы больше, можно было бы отследить окно конверсии, кластеризовать пользователей и понять, репрезентативны ли выборки.

Шаг 5. Подсчет ключевых метрик. Оценка статистической значимости полученных результатов. Оценка ошибки первого и второго рода

Campaign Name	Spend [USD]	# of Website Clicks*	actions*	# of Purchase*	CR*	CPO*
Control Campaign	66818	154303	173649	15161	112,54	4,41
Test Campaign	74595	175107	165704	14869	94,63	5,02

*статистически незначимы

Вероятность отклонить нулевую гипотезу при условии, что она верна, была сформулирована в дизайне эксперимента и составляет 5%.

Рассчитаем мощность при использовании критерия хи-квадрат. Для всех метрик, кроме CPO, мощность критерия превышает 80% (в районе единицы). Это говорит о том, что для всех метрик, кроме CPO, с вероятностью менее 1% мы бы не увидели статистически значимые различия там, где они есть. 80% - это стандартное значение мощности.

3. Выводы и рекомендации

Таким образом, на основании проведенного A/B теста можно сделать следующие выводы:

- Существует статистически значимая разница в затратах на рекламу между тестовой группой, в которой использовалась модель расчета ставки «Средняя ставка», и контрольной, где использовалась модель расчета «Максимальная ставка». Итоговые затраты на рекламу в тестовой группе составили 74595 долларов, а в контрольной 66818 долларов.
- Для остальных ключевых метрик (количество кликов, количество приобретений пользователями товаров, количество действий, конверсия, CPO) разница в показателях между тестовой и контрольной группой оказалась статистически незначимой.
- Для метрики CPO с вероятностью более чем 65% мы бы не увидели статистически значимые различия там, где они есть.

Так, можно было бы сделать общий вывод о том, что лучше использовать модель расчета ставки «Максимальная ставка», потому что затраты на рекламу меньше, а эффект от рекламной компании тот же. Однако, проведение бутстрапа заставило усомниться в правильности разделения пользователей на тестовую и контрольную группу. Мы рекомендуем провести повторный A/B тест на большее количество дней, на датасете, в котором была бы информация (целевые действия, время и дата посещения, личные характеристики (пол, возраст, география) и др.) по каждому пользователю. Так, данных было бы больше, можно было бы отследить окно конверсии, кластеризовать пользователей и понять, репрезентативны ли выборки.