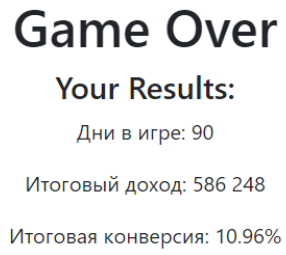
ДЗ по А/Б тестированию. Гершбаум С.Д.

Для начала приложу несколько пробных вариантов решений.

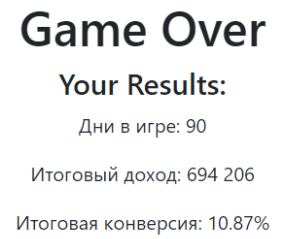
Решение 1.

Релизил только при условии, что конверсия в тестовой группе больше, чем в контрольной и $p < 0.05$ (изменения не случайны). Не релизил раньше, чем через 5 дней.



Решение 2.

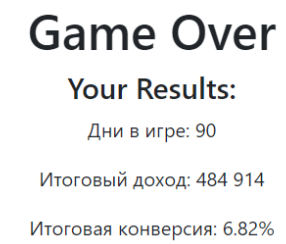
Будем релизить **только** по оценке конверсии (тест $>$ контроль). Ждем также не менее 5 дней.



Конечно, так делать не нужно. Главный показатель, на который мы смотрим – это p-value. Именно на основании него мы принимаем решение, принимается ли нулевая гипотеза H_0 или отвергается в пользу альтернативной H_A.

Решение 3.

Релизил только при условии, что конверсия в тестовой группе больше, чем в контрольной и $p > 0.05$. Не релизил раньше, чем через 5 дней.

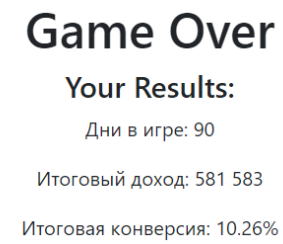


Вот и доказательство тому, что необходимо смотреть на p-value. Мы видели изменения там, где их на самом деле не было. А в некоторых случаях, видимо,

попали случайно (судя по тому, что итоговый доход $> 450\,000$ и конверсия $> 5\%$.

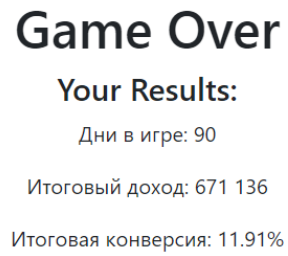
Решение 4.

Релизил только при условии, что конверсия в тестовой группе больше, чем в контрольной и $p < 0.05$ (изменения статистически значимые). Не релизил раньше, чем через 5 дней. Начинаем с мелких доработок, и по возрастающей. Рекомендации, поиск и реклама – приоритет. Интуитивно кажется, что они могут сильнее повлиять на продажи.



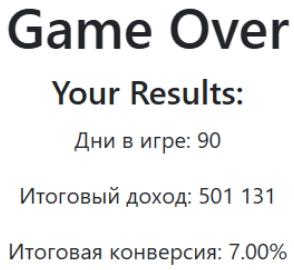
Решение 5.

Конверсия в тестовой больше, $p < 0.05$. Старался не держать более 3 активных тестов параллельно. Начинал с мелких доработок и небольших изменений функционала, был упор на рекламу и рекомендации (по возможности).



Решение 6.

Попробуем начать с больших релизов, и сделаем упор на все, кроме бэкэнда, пока он полностью не загрузит бэклог.



Решение 7.

В итоговом решении доход получился более 700 тысяч. В этот раз играл на основании наблюдений, расписанных ниже.

1. По опыту пройденных игр, для бэкэнда достаточно 2-3 дней, чтобы понять, что изменения действительно статистически значимы. Для рекламы, дизайна и рекомендаций достаточно 3-5 дней. В случае поиска почти всегда сразу понятно, что значимых изменений нет, поскольку с первого же дня и далее $p \gg 0.05$ без изменений.

2. Если активных экспериментов больше по крайней мере 3-х, то достижение MDE происходит медленнее. Поэтому большее количество тестов, проводимых одновременно, я старался не допускать.

3. Как правило, если конверсия в тестовой группе изначально значительно меньше, чем в контрольной, а если еще к тому же нулевая гипотеза уверенно отвергается ($p \ll 0.05$), то спустя дни это явление не изменится (можно сразу останавливать тест, ведь нас интересует увеличение конверсии).

4. Часто бывает так, что если подождать несколько дней, то p-value устаканивается к 0 (при «нужном» варианте конверсии). Это может зависеть от количества активных экспериментов (чем меньше активных экспериментов, тем быстрее устаканивается). Однако остановка теста не должна определяться тем, когда p-value перестает колебаться, поскольку его поведение может быть непредсказуемым. Для этого используется MDE – граничное значение эффекта (в нашем случае – увеличение конверсии), ради которого имеет смысл вводить какие-либо изменения. Т.е. хотим отлавливать с заявленной ошибкой I рода различия между тестовой и контрольной группой в $n\%$ (условно). Но поскольку это игра, с параметрами и ограниченным временем (в смысле 90 дней) не разыграешься, то за условный MDE я взял разницу $< 1\%$, чтобы как можно быстрее увеличить доход и конверсию.

5. Часто бывает так (реклама, дизайн), что на 2-4 дни при большей конверсии в тестовой группе p-значение колеблется у 0.05 (у бэкэнда и рекомендаций изменения почти всегда статистически значимы, однако прокрас разный; а у поиска чаще всего изменения статистически не значимы). Это может служить поводом для того, чтобы 1 раз взять побольше наблюдений (что происходит автоматически при переходе на следующий день) и провести тест еще 1 день (но не более), иначе мы будем растить ошибку I рода. Если из раза в раза пробовать, потому что «не прокрасилось», то далее с такой же вероятностью 5% может случиться ошибка I рода, и вероятность того, что у нас случится ошибка I рода хотя бы раз будет уже выше 5%. И чем больше мы будем подгонять, тем больше мы будем копить ошибку I рода. Поэтому если такое

случалось, я старался долго этот тест не гнать и быть «правильным» аналитиком.

6. Бэкэнд – топ-1 по статистически значимым изменениям (чуть ли не в 100% тестов получается то, чего мы так ждем). После него в топе идут реклама и дизайн. Поиск и рекомендации почти никогда не отправляются в релиз. Выиграть в итоговой выручке можно тогда, когда в бэклоге на старте игры сразу есть актуальные задачи для команды бэкэнда независимо от типа задачи (но, к сожалению, это рандом).

7. По логике, надеясь на статистически значимые изменения (а именно зеленый прокрас, конверсия в тестовой группе больше, чем в контрольной) и максимальную выручку по итогу 90 дней, логично начинать с задач, на подготовку которых уходит меньше всего времени (мелкая доработка, небольшое изменение функционала), чтобы прирост дохода (увеличение конверсии) начался как можно раньше (если в надежде ждать зеленый прокрас сразу же). Но самый удачный вариант – если на старте есть задача от команды бэкэнда, о чем сказано в предыдущем пункте.