

Бизнес-оценка рекомендательных систем

Дмитрий Шипилов

ML Engineer, Uzum Market

Skillbox

Планирование A/B-тестирования

Цели темы

- 1 Научиться планировать A/B-тест
- 2 Изучить, как рассчитывать итог A/B-теста
- 3 Понять, зачем нужны A/A-тесты

A/B-тест

A/B-тест (A/B-эксперимент) — это исследование, суть которого заключается в сравнении поведения пользователей в нескольких различных группах.

Сравнение двух версий

Конверсия старой версии: $(30/920) \times 100 = 3,2$

Конверсия новой версии: $(35/900) \times 100 = 3,8$

	Старая версия	Новая версия
Количество просмотров	920	900
Количество кликов	30	35
Конверсия, %	3.2	3.8

Этапы тестирования

- 1 Предварительная проверка
- 2 Выбор метрики
- 3 Нулевая гипотеза и уровень значимости
- 4 Определение размера эффекта
- 5 Подсчёт размера групп и длительности теста
- 6 Выбор критерия и подведение результатов

Предварительная проверка модели

Разметка — это процесс оценки, отображающий, насколько эти предсказания выглядят осмысленно и релевантно для данного юзера.

- Прогнозы должны отличаться для каждого из пользователей
- Имеет смысл провести разметку на части прогнозов

Выбор метрики

Защитные метрики — это самые важные метрики проекта, которые не должны стать хуже в рамках проводимого теста. Например, ARPU — средняя прибыль проекта с одного пользователя (Average Revenue Per User).

Метрика:

- должна отражать бизнес ценность
- вы должны влиять на метрику в тесте

Нулевая гипотеза

Гипотеза — предположение об отличии между группами.

Нулевая гипотеза (или H_0) — это консервативное предположение о том, что между нашими группами нет никаких различий, например: «конверсия новой рекомендательной системы не отличается от прошлой».

Альтернативная гипотеза (H_1) утверждает обратное: «конверсия новой рекомендательной системы отличается от прошлой (не обязательно в хорошую сторону)».

Уровень значимости α и мощность

Уровень значимости α — максимальная вероятность отклонить верную нулевую гипотезу.

Чем ниже значение α , тем больше наблюдений нужно будет в каждой из групп.

Мощность — способность засечь разницу между группами.

Часто используемые значения:

- α — 5 % (0,05)
- мощность — 80 % (0,8)

Ошибки I и II рода

Ошибка I рода — отклонить нулевую гипотезу, хотя она была верной, ограничивая с помощью уровня значимости.

Ошибка II рода — принять неверную нулевую гипотезу, ограничивая с помощью мощности.

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка второго рода)
	H_1	H_0 неверно отвергнута (Ошибка первого рода)	H_0 верно отвергнута

Размер эффекта

Ожидаемый размер эффекта можно предположить исходя из:

- исторических данных
- опыта
- офлайн-метрики

Если офлайн – метрика (например, nDCG) выросла несильно, то и онлайн-метрика, скорее всего, тоже вырастет не особо сильно.

Эффект бывает абсолютным и относительным:

конверсия новой системы 4 % вместо 3 % значит абсолютный рост в 1 %, а относительный — почти 33 %.

Длительность теста и размер групп

На основе уровня значимости и ожидаемого размера эффекта можно сказать, сколько пользователей необходимо в группе A — тестовой — и группе B — контрольной.






На основе этого можно подсчитать, сколько времени необходимо крутить тест до того, как в каждой группе будет нужное количество юзеров.

Калькуляторы:

- [abtestguide](#)
- [evanmiller](#)
- [select-statistics](#)

Длительность теста и размер групп

Calculator

What confidence level do you need? <small>Typical choices are 90%, 95% or 99%</small>	<input type="text" value="95"/> %	
What power do you need? <small>A common choice is 80%</small>	<input type="text" value="80"/> %	
What do you believe the likely sample proportion in group 1 to be?	<input type="text" value="3"/> %	
What do you believe the likely sample proportion in group 2 to be?	<input type="text" value="4"/> %	
Your recommended sample size is		5298 

Подведение итогов теста: критерии

Статистический критерий — математический инструмент, который принимает на вход данные о ваших распределениях и на их основе подсчитывает специальные статистики.

Реализованы в пакетах SciPy и Statsmodels.

Например, для конверсий подойдёт `test_proportions_2indep` из Statsmodels.

Подведение итогов теста и p-value

P-value — это вероятность получить такое или ещё более экстремальное распределение статистики, если нулевая гипотеза верна, рассчитывается с помощью статистического критерия.

Сравнивается с выбранным уровнем значимости α .

Если $p_value < \alpha$, то можете отклонить нулевую гипотезу.

Если $p_value > \alpha$, то данные не говорят о том, что нулевую гипотезу можно отклонить.

Множественное тестирование

Если гипотез несколько, например, у вас больше двух групп, то один из способов избежать ошибки — поправка Бонферрони, т. е. нормирование выбранного уровня значимости α на фактическое количество групп N .

Множественное тестирование

N — общее количество групп (например, 3)

α — выбранный уровень значимости (0,05)

$\alpha_{\text{new}} = \alpha/N$ новый уровень значимости

$\alpha_{\text{new}} = 0,05/3 = 0,016$

A/A-тесты

Попытка найти различия, когда в группах их фактически нет.

Помогает понять, что с вашей системой всё хорошо.

Выводы

- 1 Узнали, как планировать A/B-тест
- 2 Рассмотрели множественное тестирование
- 3 Поняли, зачем нужны A/A-тесты