Планирование дизайна A/B

Юрий Котов
Senior Data Engineer
Tinkoff

План лекции

- Разбиение на группы
- Дизайн эксперимента
- Бутстреп и доверительные интервалы

План лекции

- Разбиение на группы
- Дизайн эксперимента
- Бутстреп и доверительные интервалы

Как разбить пользователей на группы?

- По идентификаторам пользователя (уникальные id, логины, почта, номер телефона и т.д)
- По файлам cookie (на разных устройствах отличается, обновляется при закрытии браузера и т.д), id устройства

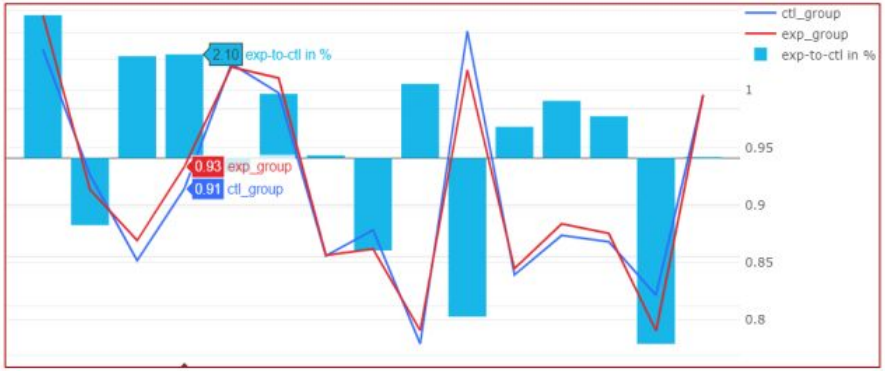
После разбиения надо проверить его качество, как?

Как разбить пользователей на группы?

- По идентификаторам пользователя (уникальные id, логины, почта, номер телефона и т.д)
- По файлам cookie (на разных устройствах отличается, обновляется при закрытии браузера и т.д), id устройства

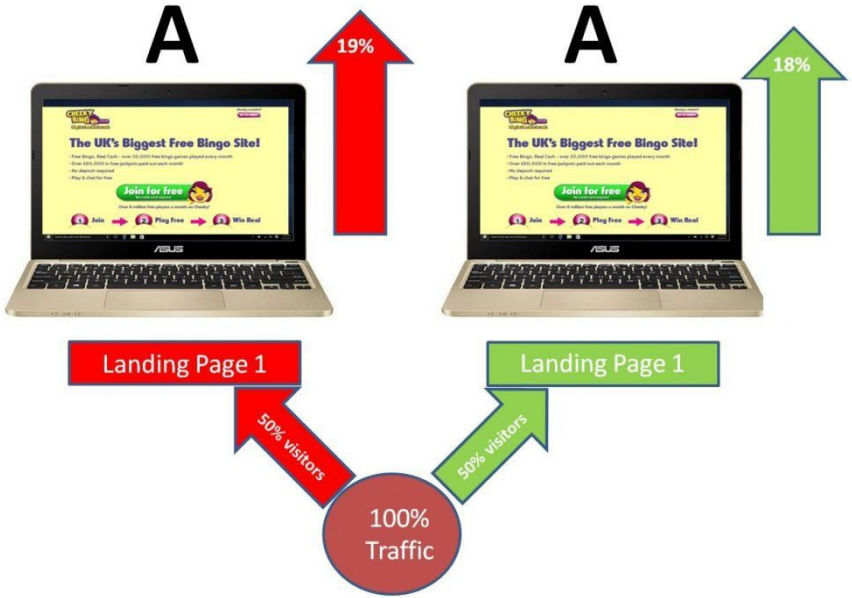
После разбиения надо проверить его качество, как?

Поможет АА-тест



AA тест: что это такое и зачем он нужен

Один из способов проверить равномерность разбивки пользователей, убедиться, что нет статзначимых изменений там, где их быть не должно. И выбранный критерий для оценки результата корректен и обладает достаточной мощностью.



Рандомное разбиение

Для разбиения используют hash от id пользователя с солью.

Часто используемые hash-функции:

- SHA-2
- Cityhash
- MD5

Соль позволяет избавится от зависимости от id-шника => при разных вариантах соли можно получать разные разбивки

Пример: берём id пользователя, добавляем к нему «соль», хэшируем, приводим к виду целого числа и берем остаток от деления на 2 (если группы 2)

Но могут быть сложности

Выбирая из генеральной совокупности часть пользователей для теста, мы рискуем случайно отобрать нерепрезентативную подвыборку или же поделить группы неравномерно в каком-то аспекте (особенно при малых выборках)

Как можно этого избежать?

Но могут быть сложности

Выбирая из генеральной совокупности часть пользователей для теста, мы рискуем случайно отобрать нерепрезентативную подвыборку или же поделить группы неравномерно в каком-то аспекте (особенно при малых выборках)

Как можно этого избежать?

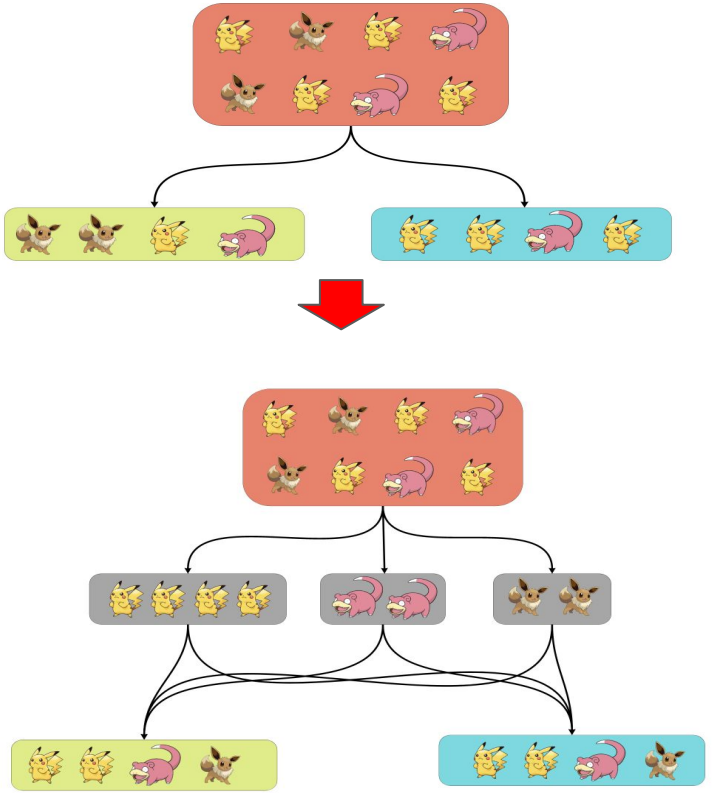
Для балансировки групп можно выделить важные для нас признаки, по которым будем проверять однородность.

Например, возраст или доход, средний чек или регион и т.д

Другие подходы к подбору групп

Стратификация

По интересующему нас параметру разбиваем генеральную совокупность на страты и из каждой страты в равном количестве набираем объекты в группы для теста. Число элементов из каждой страты в группах должно быть пропорционально размеру этой страты в исходной совокупности



Стратификация: как работает

$$Y_{strat} = \sum_{k=1}^{K} w_k Y_k$$

w_k – вес группы k (например, вероятность попадания в эту группу),
Y_k – метрика в группе k
Y- целевая метрика

$$Var_{strat} = \frac{1}{N_t} \times \sum_{k=1}^{K} p_k \sigma_k^2$$

$$Var_{samp} = Var_{strat} + \frac{1}{N_t} \times \sum_{k=1}^{K} p_k (\mu - \mu_k)^2$$

K — количество подгрупп
p_k — вероятность случайного наблюдения в изначальной выборке оказаться наблюдением типа k
σ_k — стандартное отклонение метрики в подгруппе k.
μ_k — среднее метрики в подгруппе k
μ — среднее метрики на всей совокупности наблюдений

Дисперсия стратифицированной выборки состоит из взвешенных дисперсий внутри страт. А дисперсия при рандомном разбиении состоит из дисперсии стратифицированной выборки и взвешенной дисперсии между стратами

Стратификация: вывод

- Метод стратификации обеспечивает объективную оценку эффекта лечения и эффективно устраняет дисперсию между подгруппами
- Метод хорошо подходит для маленьких или средних выборок
- Дает заметный эффект, если текущее сэмплирование групп смещено по каким-либо признакам
- Однако на практике обычно очень сложно реализовать стратифицированную выборку перед экспериментом

Другие подходы к подбору групп

Подбор групп по критериям однородности

Выбираем целевую метрику и/или любой другой исторический признак, по которым хотим найти похожие группы. Выбранным критерием для разбивки (Манн-Уитни/Колмогоров-Смирнов) ищем разбиения, для которых выборки идентичны

Другие подходы к подбору групп

Подбор групп по прогнозам целевой метрики

Используя линейную регрессию, прогнозируем целевую метрику на период теста. По прогнозным значениям разбиваем на группы (например, с помощью критериев однородности)

https://www.tripadvisor.com/engineering/reducing-a-b-test-measurement-variance by-30/

Другие подходы к подбору групп

Подбор групп по индексу Аткинсона

Экономический индекс, используемый для оценки социального неравенства. Можно использовать для подбора сбалансированных групп

https://arxiv.org/abs/2002.05819

План лекции

- Разбиение на группы
- Дизайн эксперимента
- Бутстреп и доверительные интервалы

Подготовка к тесту

- Сформулировать гипотезу + выбрать целевую метрику
- Определиться с разбивкой на группы
- Провести дизайн эксперимента:
  1. Выбор статкритерия
  2. Расчет размера выборки/длительности теста
  3. Расчет минимального эффекта (учитываем и бизнес-смысл, и вычисляемую величину)
  4. Выбор уровней ошибок 1 и 2 рода

Дизайн теста

Перед запуском теста необходимо определиться с основными параметрами: эффект, продолжительность, размер выборки, ошибки 1 и 2 рода

Как величина ошибок влияет на продолжительность теста?

- Чем выше мощность, тем более низкий эффект мы сможем статзначимо поймать, но тогда нужно больше данных => дольше тест
- Чем ниже мощность, тем реже мы сможем поймать эффекты со статзначимостью, особенно низкие эффекты, но тем меньше наблюдений нам требуется => короче тест

MDE и минимальный размер выборки

MDE - это показатель, который относится к наименьшему эффекту, который на практике имел бы определенный уровень мощности для определенного уровня статистической значимости с учетом конкретного статистического теста.

Минимальный размер выборки - минимальное количество наблюдений в контрольной и экспериментальной группах, которое необходимо набрать с учетом определенной мощности теста (1-бета), уровня значимости (альфа) для обнаружения минимального эффекта (MDE).

Продолжительность теста зависит от минимального размера выборки

N – размер выборки, σ – стандартное отклонение, α и β – ошибки 1 и 2 рода, соответственно

$$MDE = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right) * \sqrt{\left(\sigma^2_{control} + \sigma^2_{experiment}\right)}}{\sqrt{N}}$$

$$N = \frac{\left(\sigma^2_{control} + \sigma^2_{experiment}\right) * \left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2}{MDE^2}$$

MDE и минимальный размер выборки: пример

N = 200000 пользователей
σ = 0,49 (допустим, что в группах равные стандартные отклонения)
α = 0,05 и β = 0,2
Z = 1,96 (для 95% значимости)

$$MDE = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right) * \sqrt{\left(\sigma^2_{control} + \sigma^2_{experiment}\right)}}{\sqrt{N}}$$

$$MDE = \frac{(1{,}96 + 0{,}84) * \sqrt{2 * 0{,}49^2}}{\sqrt{200000}}$$

MDE = 0,0043, но это не разница между группами!

Чтобы получить прирост в процентах, который мы хотим увидеть между группами, надо MDE * 100% $/mean_{control}$

Например, $mean_{control}$ = 0,453, тогда мы сможем зафиксировать эффект в ≈ 0,95%

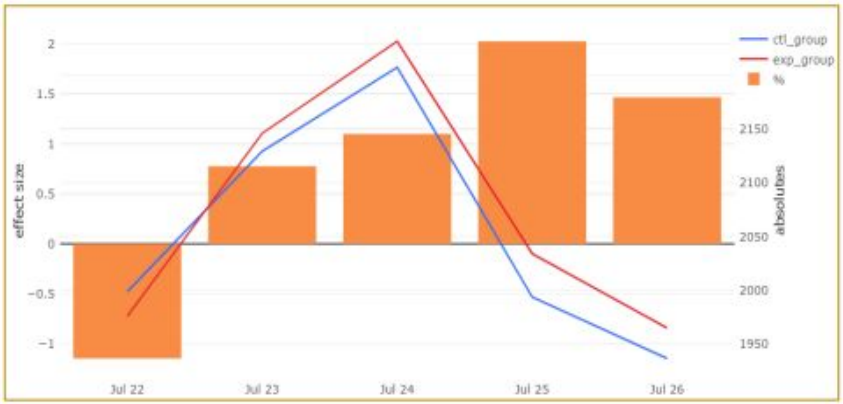
А если мы не хотим ждать…

На дизайне выяснилось, что для «отлова» эффекта в 1% нам понадобится ждать 3 недели, иначе наблюдений не хватает.

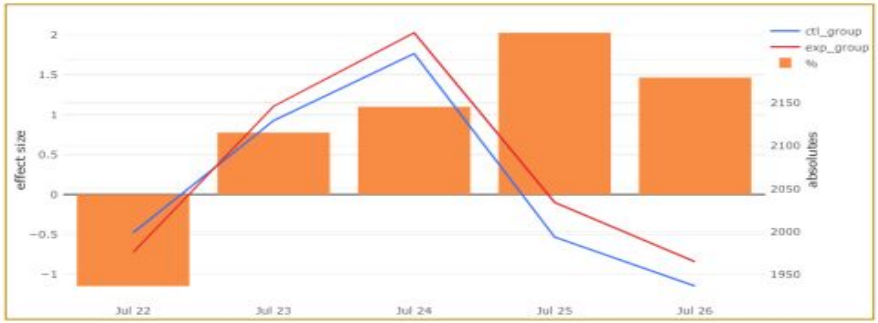
Но у нас есть мониторинг с подневными данными по ходу эксперимента, где отражается эффект, набранный к определенному моменту + рассчитывается каждый день p-value

И вдруг мы видим, что произошел «прокрас». Можно завершить тест раньше?

Что мы можем сказать об этих тестах?

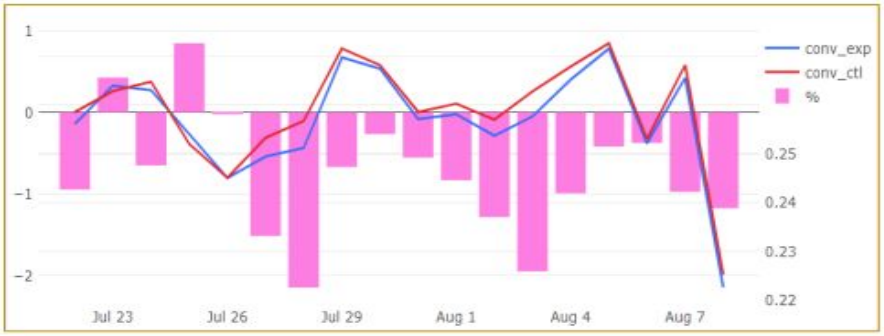


Что мы можем сказать об этих тестах?



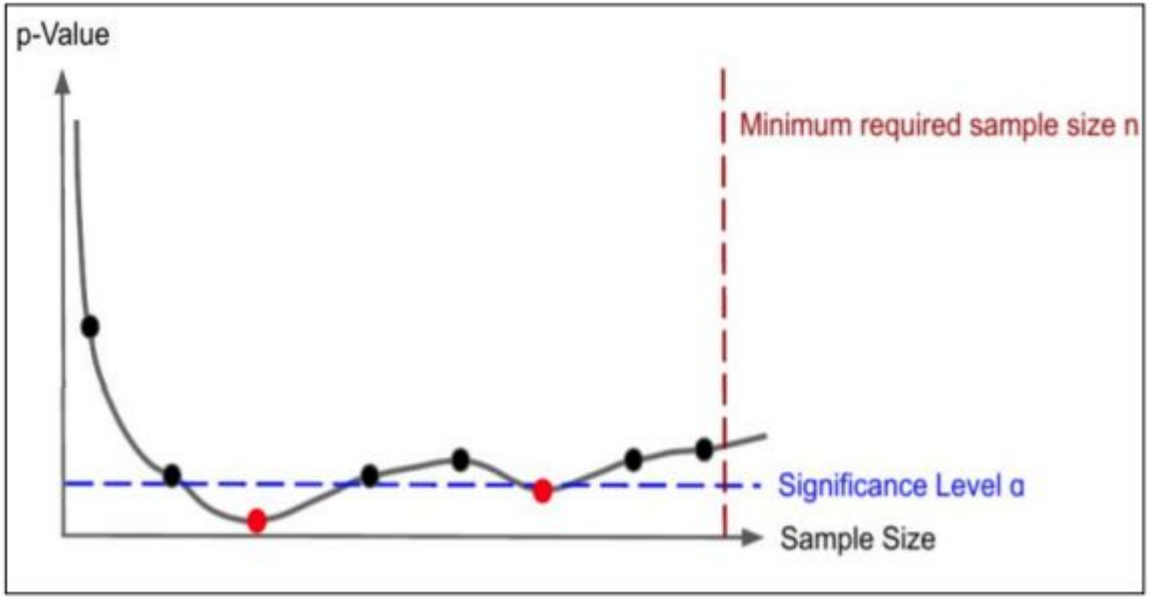
Нет статзначимого эффекта

Статзначимый отрицательный эффект



Проблема подглядывания

Происходят ложные прокрасы метрики => растет вероятность ошибки 1 рода



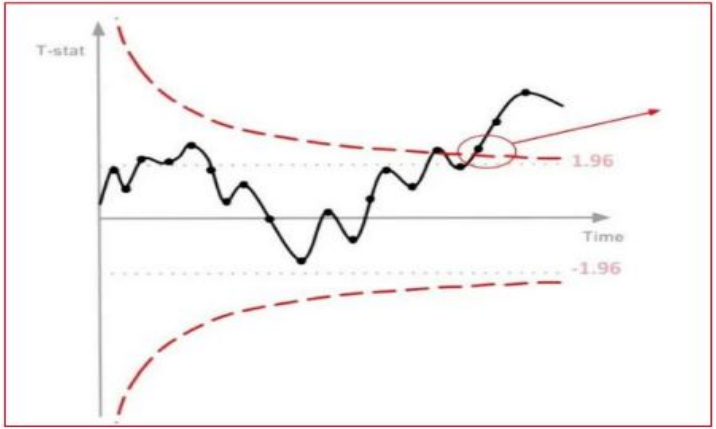
Проблема подглядывания: решение

Вариант 1: держим тест до конца

- Делаем честный дизайн => представляем, чего ожидать
- Строим мониторинги и наблюдаем за ходом эксперимента (но не делаем выводы!)

Вариант 2: используем альтернативные подходы

- Sequential testing aka Последовательный анализ



А если больше 2-х групп и/или 1-й метрики…

Например, в нашем тесте не 2 группы (контроль и эксперимент), а 3: два эксперимента и один контроль

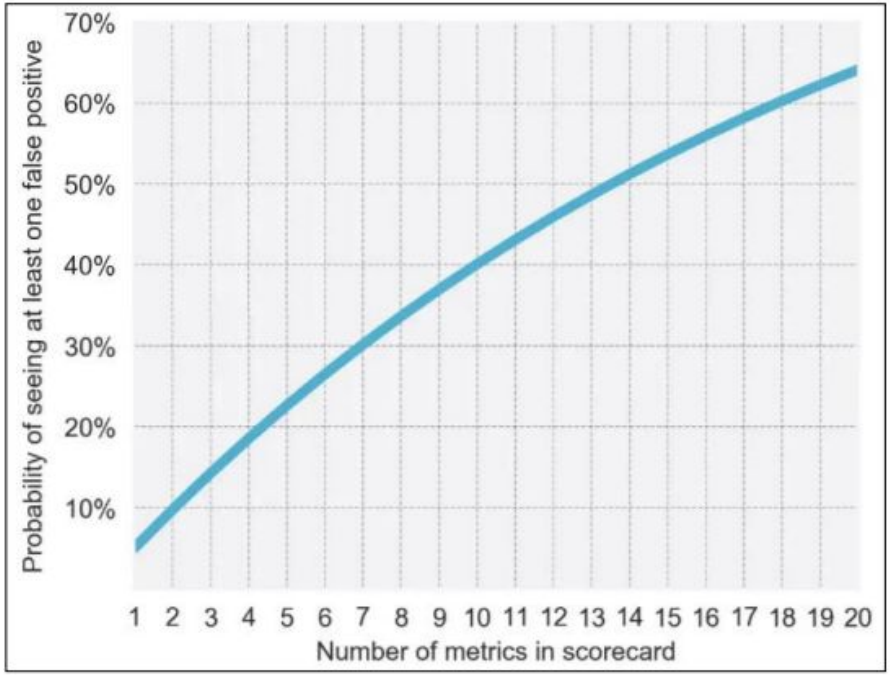
Или же группы 2, а целевых метрик тоже 2, а не 1

Тогда мы сравниваем уже не один раз контроль с экспериментом, а производим несколько попарных сравнений:

- Для случая с тремя группами и одной метрикой – 3, если сравниваем все со всеми, или 2, если только эксперименты с контролем
- Для случая с двумя группами и двумя метриками – 2 (по одному сравнению на каждую метрику)

Одновременное проведение нескольких тестов

- Коэффициент ложноположительных результатов для одного теста: 0,05
- Вероятность не совершить ошибку типа I за один тест: 1 — 0,05
- Вероятность не совершить ошибку типа I за 10 тестов: $(1 — 0,05)^{10}$
- Вероятность совершения ошибки типа 1 хотя бы один раз за 10 тестов: $1 — (1 — 0,05)^{10} = 0,40$, или 40%



Одновременное проведение нескольких тестов: решение

Введение поправок на множественную проверку гипотез

Бонферрони: необходимо отклонить те гипотезы, для которых p-value <α/m (m-количество попарных сравнений)

- Плюс – легко реализуема
- Минус – сильно «съедает» мощность

Одновременное проведение нескольких тестов: решение

Введение поправок на множественную проверку гипотез

Холма-Бонферрони: считаем p-value и упорядочиваем их по возрастанию. Далее сравниваем по очереди наши p-value с α/(m −k+1) , где k от 1 до m, m − количество проверяемых гипотез.
Если p-value >= α/(m −k+1) , то останавливаемся и принимаем текущую и последующие нулевые гипотезы

- Плюс – мощнее Бонферрони
- Минус – чуть сложнее в вычислениях

https://habr.com/ru/articles/772940/

План лекции

- Разбиение на группы
- Дизайн эксперимента
- Бутстреп и доверительные интервалы

Бутстрэп

Суть метода: путем многократного повторного извлечения наблюдений из исходной выборки мы можем построить эмпирическое распределение, которое поможет апроксимировать исходное неизвестное распределение необходимой нам статистики

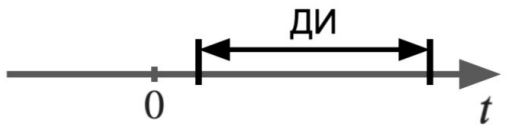
- Имеет в основе идею ресэмплинга (многократную генерацию повторных выборок на основе одних и тех же данных)
- Можно оценивать средние/дисперсию/квантили, строить доверительные интервалы, тестировать гипотезы
- Выборка может быть небольшой, но отражающей генеральную совокупность
- Случайные величины в выборке независимы

Бутстрэп

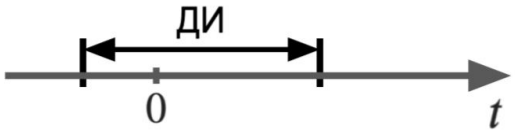
Pipeline применения bootstrap для A/B-тестирования:

1. Разбиваем пользователей на группы, выбираем метрику, формулируем гипотезы, проводим тест
2. В конце теста генерируем подвыборки того же размера, что исходные группы (для несмещенности оценок) контроля и эксперимента не менее 1000 раз
3. Считаем метрики для обеих групп и вычисляем их разность на каждой из этих 1000+ пар подвыборок
4. Строим доверительный интервал для разности метрик с уровнем значимости α и смотрим, содержит ли он 0
5. Если 0 попадает в доверительный интервал – принимаем H_0, в ином случае – H_1

Есть статистически значимые отличия



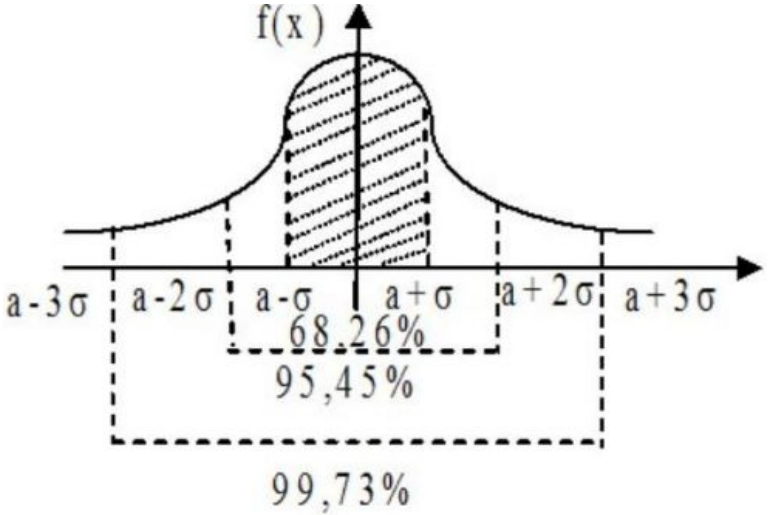
Статистически значимых отличий нет



Бутстрэп и доверительные интервалы

- Для каждой подвыборки, сгенерированной во время проведения бутстрэпа мы строим точечную оценку интересующего нас параметра
- Собрав достаточно точечных оценок, мы получаем их распределение
- Для этого распределения можно построить доверительный интервал с интересующими нас границами

Правило 3-х сигм



Для случайных величин, распределенных нормально:

- с вероятностью ≈ 95% эта случайная величина принимает значения в интервале $\mu \pm 2\sigma$
- с вероятностью 99.72% эта случайная величина принимает значения в интервале $\mu \pm 3\sigma$

$$P(|X-a|<\sigma) = P(a-\sigma<X<\sigma+a) = 0,6826$$
$$P(|X-a|<2\sigma) = P(a-2\sigma<X<2\sigma+a) = 0,9545$$
$$P(|X-a|<3\sigma) = P(a-3\sigma<X<3\sigma+a) = 0,9973$$

Доверительные интервалы

Доверительный интервал покрывает неизвестный параметр с заданной надежностью

Интервал $[\theta_L; \theta_U]$ называется доверительным интервалом для параметра θ, с уровнем доверия $1 - \alpha$, если при бесконечном повторении эксперимента в $100 \cdot (1 - \alpha)$ процентах случаев этот интервал будет включать истинное значение параметра θ

α - уровень значимости - вероятность, с которой значение параметра не попадает в доверительный интервал

Обычно уровень значимости выбирают 0.01 или 0.05 (реже 0.1, т.е. 10%), что соответствует уровню доверия 0.99 или 0.95, соответственно

При расчете доверительного интервала мы можем задать вероятность попадания фактических значений в заданные границы прогноза.

Границы самого интервала представляют собой случайные величины, которые мы пытаемся найти по выборке

Доверительный интервал для среднего

n – объем случайной выборки, $\bar{x}$ - среднее значение выборки, уровень доверия 1 – α, σ^2 - известная дисперсия выборки, s – несмещенное выборочное стандартное отклонение
Квантиль нормального распределения уровня 1 – α/2 это $z_{1-\frac{\alpha}{2}}$

Квантиль распределения Стьюдента – это $t_{1-\frac{\alpha}{2}}$

Дисперсия известна

$$P\left(\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Дисперсия неизвестна

$$P\left(\bar{x} - t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n-1}}\right) = 1 - \alpha$$

Доверительный интервал для среднего имеет вид $(\bar{x} - \Delta; \bar{x} + \Delta)$, где Δ - это точность интервальной оценки

Доверительный интервал для разности средних (независимые выборки)

Дисперсии известны

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(\mu, \sigma^2)$$

Можем использовать квантиль нормального распределения $z_{1-\frac{\alpha}{2}}$

Дисперсии неизвестны (и неравны)

$$\frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{s^2}{n_x} + \frac{s^2}{n_y}}} \sim t(v)$$

Если дисперсии равны, то $\sim t(n_x + n_y - 2)$
Используем распределение Стьюдента

$$\bar{x}_1 - \bar{x}_2 - Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{s_x^4}{n_x^2(n_x - 1)} + \frac{s_y^4}{n_y^2(n_y - 1)}}$$

Если неравны, то модификация t-распределения Уэлша

Доверительный интервал для дисперсии

n – объем случайной выборки, уровень доверия 1 – α, s – несмещенное выборочное стандартное отклонение
Квантили распределения хи-квадрат χ_n^2 для 1 – α/2 и α/2 ищутся в таблице
(с n степенями свободы, когда матожидание известно, и с n - 1 степенями свободы при неизвестном матожидании

Математическое ожидание известно

$$P\left(\frac{n \cdot s^2}{\chi_n^2\left(1-\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{n \cdot s^2}{\chi_n^2\left(\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

Математическое ожидание неизвестно

$$P\left(\frac{(n-1) \cdot s^2}{\chi_{n-1}^2\left(1-\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-1) \cdot s^2}{\chi_{n-1}^2\left(\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

Доверительный интервал для отношения дисперсий

n, m – объемы случайных выборок, s_n^2 – квадрат несмещенного выборочного стандартного отклонения выборки n, s_m^2 - то же самое для выборки m

Т.к. дисперсия представляет собой квадрат, вместо разности вычисляем отношение

$$\frac{(m-1)\cdot s_m^2}{\sigma_m^2} \sim \chi_{m-1}^2$$

$$\frac{(n-1)\cdot s_n^2}{\sigma_n^2} \sim \chi_{n-1}^2$$

$$\Rightarrow \quad \frac{\frac{(n-1)\cdot s_n^2}{\sigma_n^2}}{n-1} \Big/ \frac{\frac{(m-1)\cdot s_m^2}{\sigma_m^2}}{m-1} = \frac{\chi_{n-1}^2}{n-1} \Big/ \frac{\chi_{m-1}^2}{m-1}$$

Т.к. отношение величин, имеющих хи-квадрат распределение и скорректированных на количество степеней свободы, подчиняется распределению Фишера:

$$\frac{s_m^2}{s_n^2} \cdot F_{n-1,m-1}\left(\frac{\alpha}{2}\right) \leq \frac{\sigma_m^2}{\sigma_n^2} \leq \frac{s_m^2}{s_n^2} \cdot F_{n-1,m-1}\left(1-\frac{\alpha}{2}\right)$$

Доверительные интервалы: трактовка

Вопрос: что означает фраза "истинное значение параметра лежит в границах [θ1 ; θ2] с вероятностью 0.95"?

Доверительные интервалы: трактовка

Вопрос: что означает фраза "истинное значение параметра лежит в границах [θ1 ; θ2] с вероятностью 0.95"?

При проведении серии повторяющихся оценок параметра (в одинаковых условиях выбора наблюдений и оценки параметра) в 95% случаев доверительный интервал для оцениваемого параметра будет содержать его (параметра) истинное значение. В 5% случаев произойдет ошибка, т.е. интервал не будет включать истинное значение оцениваемого параметра

Выводы

1. Метрику, разбивку и выборку можно проверить на AA-тесте
2. Разбиение на группы должно быть случайным (или же заранее запланированным в случае стратификации)
3. Перед началом теста необходимо сформулировать нулевую и альтернативную гипотезы, а также провести дизайн эксперимента
4. При дизайне стоит учесть ошибки 1 и 2 рода и определить продолжительность эксперимента в зависимости от размера выборки и минимального детектируемого эффекта
5. Не прекращайте тест раньше времени + не считайте p-value каждый день
6. Выбирайте 1 целевую метрику/ тестируйте 1 гипотезу или вводите поправки на множественную проверку гипотез
7. Если наша метрика специфична и известные нам критерии не подходят для ее оценки, можно попробовать применить бутстрэп (с соблюдением всех ограничений)
8. Иногда бизнесу может быть важно знать не только точечную оценку, а интервал, поэтому хорошей практикой станет построение доверительных интервалов

Спасибо за внимание!