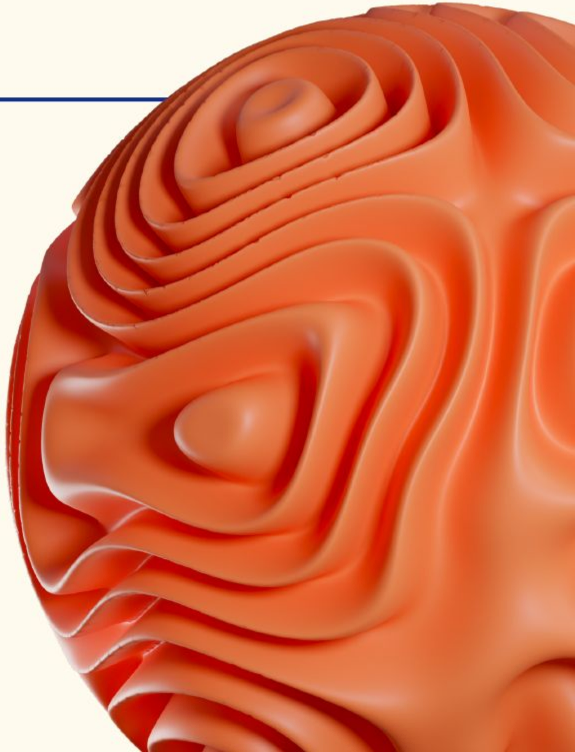
Модели ML в production

Лекция № 12
“AB-тесты: advanced”



Жарова Мария Александровна
DS WB-tech, Math&Python&DS lecturer
t.me/data_easy

План занятия

1. Доверительный интервал
2. Практика: часть 1
3. Bootstrap vs стат. критерии
4. Способы уменьшения дисперсии:
    - Стратификация
    - CUPED
5. Практика: часть 2

Ошибка I и II рода наглядно

Гипотеза: человек не беременный.

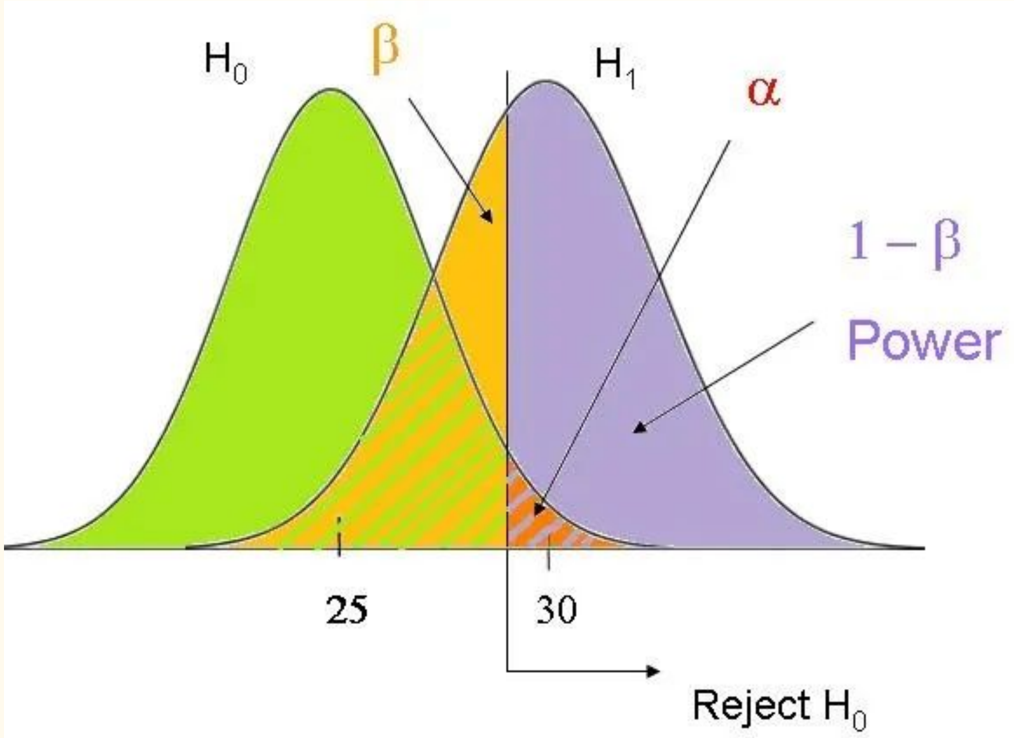
Ошибка I рода: гипотеза верна, но мы её отвергаем.

Ошибка II рода: гипотеза неверна, но мы её принимаем.

Статистический критерий

— это математическое правило, позволяющее по реализациям выборок отвергнуть или не отвергнуть *нулевую гипотезу* с *заданным уровнем значимости*.

В ходе прим. стат. критерия получаем значение статистики ⇒ p-value



1. Доверительный интервал

Доверительный интервал

Доверительный интервал показывает **диапазон в котором лежит среднее значение выборки с вероятностью (1 - a)**:

$$P\left(\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\delta}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\delta}{\sqrt{n}}\right)$$

- Чем больше число наблюдений, тем меньше доверительный интервал, (логично, так мы можем точнее оценить средне).
- Доверительный интервал и правило 2 сигм - разные вещи, которые часто путают!
  Д.И. показывает диапазон среднего случайной величины, правило 2 сигм показывает диапазон с 2,5 до 97,5 квантиля случайной величины.
- Доверительный интервал при увеличении выборки уменьшается, квантили не меняются!

Интерпретация

Мы не говорим, что *"μ лежит внутри этого конкретного интервала с вероятностью 95%"*, а говорим, что *если бы мы многократно повторяли эксперимент и строили интервалы, то 95% из них накрывали бы μ*.

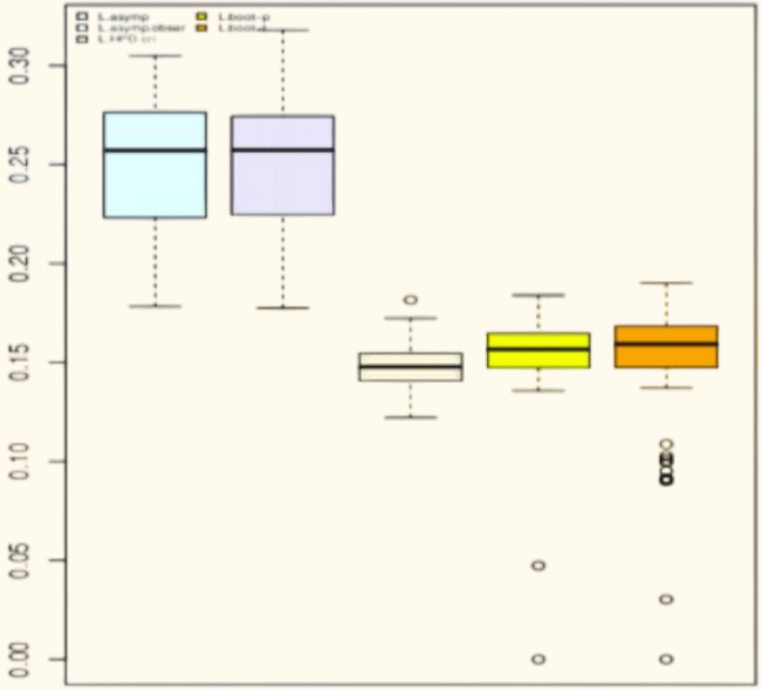
Плохой пример из жизни

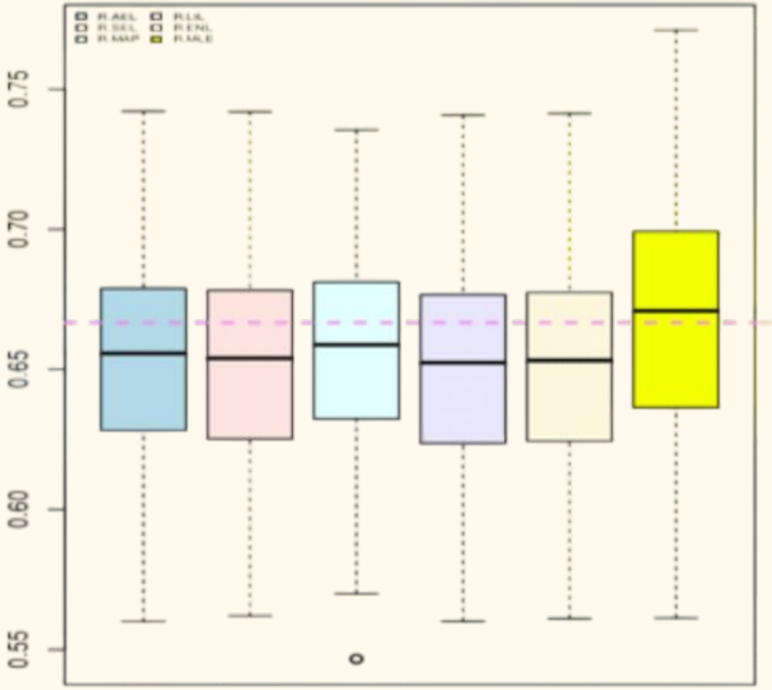
Вследствие неверного дизайна АБ-теста размеры тестируемых групп оказались недостаточного размера.

Тогда нельзя гарантировать, что статзначимость будет рассчитана корректно!

<u>На постанализе возможна ситуация:</u> *pvalue < alpha* (есть стат. значимость), но доверительные интервалы будут пересекаться ⇒ это как раз указывает на то, что средние значения в группах могут оказаться одинаковыми, несмотря на результат стат. критерия.

Сравнение дов. интервалов





2. Практика: часть 1

3. Bootstrap vs стат. критерии

Бутстрэп: идея

– это метод повторной выборки с возвращением из имеющихся данных для оценки неопределенности статистик.

Ключевая идея:

1. у нас есть одна выборка (а не теоретическое распределение), но мы хотим узнать, насколько точно мы можем оценить, например, среднее
2. мы симулируем «много выборок» из этой одной, чтобы оценить дисперсию и доверительный интервал интересующей нас статистики

Бутстрэп: когда применять

<u>Бутстрэп полезен, когда:</u>

- нет аналитической формулы для дисперсии или доверительного интервала (например, для медианы, ROC AUC, сложных метрик);
- выборка слишком мала или не имеет нормального распределения (например, данные с перекошенным распределением);
- сложно использовать стандартные статистические критерии, потому что нарушаются их предпосылки (например, неоднородность дисперсий, негауссовские данные).

<u>Минусы:</u> дольше работает:(

Бутстрэп: алгоритм

<u>Пример:</u> Рассмотрим гипотезу о равенстве математических ожиданий в двух выборках: ***m1 = m2*** против альтернативы: ***m1 != m2***, на уровне значимости 5%

<u>Алгоритм:</u>

- Есть две выборки X_1 и X_2. Генерируем ***K*** пар подвыборок размера ***n*** из них (с возвращением).
- Для каждой пары считаем разность выборочных средних $T = X_{i1} - X_{i2}$, $i = 1 \dots K$.
- Получаем набор разностей $T_1 \dots T_K$.
- Если уровень значимости – 5%, обрезаем выборку: слева – 2,5% квантилем, справа – 97,5% квантилем.
- Если 0 не входит в получившийся интервал, то гипотеза отвергается.

Стат. критерий или бутстрэп?

В большинстве случаев они работают одинаково, но есть случаи, когда можно применить только бутстрэп или только стат. критерий.

Пример № 1: тестируем новые фичи в модели и хотим понять, как меняется ROC AUC. Здесь мы не можем применить, например, тест Стьюдента, так как у нас *просто нет выборки, по которой мы сможем посчитать дисперсию разницы* ROC AUC.

Алгоритм:

- Фиксируем гиперпараметры старой и новой моделей.
- Запускаем цикл и в каждой итерации берем подвыборку из нашего датафрейма с повторениями, строим модель, считаем разницу ROC AUC новой и старой моделей.
- Получаем набор разностей ROC AUC.
- Ограничиваем набор: слева 2,5%, справа 97,5% квантилем.
- Если 0 не входит в получившийся интервал, то гипотеза отвергается, значит новые признаки круто повышают качество!

Стат. критерий или бутстрэп?

Пример № 2: сделали новую онлайн модель и скорость принятия решений крайне важна.

- В этом случае бутстрэп будет малоэффективен, так как он работает довольно долго.
- Стат. критерии (например, тест Стьюдента), напротив, будут работать моментально.

- $T = \dfrac{\overline{X}_{test} - \overline{X}_{control}}{\sqrt{\dfrac{s^2_{test}}{n_{test}} + \dfrac{s^2_{control}}{n_{control}}}}$, где

- $\overline{X}_{test}$ и $\overline{X}_{control}$ - средние значения в тестовой и контрольной группах
- s^2_{test} и $s^2_{control}$ - дисперсии в тестовой и контрольной группах
- n_{test} и $n_{control}$ - количество наблюдений в тестовой и контрольной группах

PS: напоминание про Стьюдента

- $$T = \frac{\overline{X}_{test} - \overline{X}_{control}}{\sqrt{\frac{s_{test}^2}{n_{test}} + \frac{s_{control}^2}{n_{control}}}}, \text{ где}$$

- $\overline{X}_{test}$ и $\overline{X}_{control}$ - средние значения в тестовой и контрольной группах
- s_{test}^2 и $s_{control}^2$ - дисперсии в тестовой и контрольной группах
- n_{test} и $n_{control}$ - количество наблюдений в тестовой и контрольной группах

- Проверяет гипотезу о равенстве математических ожиданий в двух выборках: $\mu_1 = \mu_2$ против альтернативы: $\mu_1 \neq \mu_2$
- Как принимать решение? Есть два варианта:
  - Если $T > T_{крит}$ или $T < -T_{крит}$ то гипотеза отклоняется, $T_{крит}$ считается в зависимости от уровня значимости α, чаще всего выбирают 0,05, $T_{крит}$ при этом равен 1,96
  - Считаем p_{value} для T, если $p_{value} < \alpha$, то гипотеза отклоняется

4. Способы уменьшения дисперсии

Вспомним MDE

$$n > \frac{\left[z_{1-\alpha} + z_{1-\beta}\right]^2 \cdot \left(\sigma_x^2 + \sigma_y^2\right)}{\varepsilon^2}$$

- *n – необходимое кол-во наблюдений в каждой группе (test и control)*
- *α – уровень значимости (обычно 0.05)*
- *β – вероятность ошибки II рода (обычно 0.2, значит 80% мощность)*
- *ε – тот самый MDE (эффект, который хотим заметить)*
- *σ_x^2, σ_y^2 – дисперсии в контрольной и тестовой группах*
- *$z_{1-\alpha}$, $z_{1-\beta}$ – квантили стандартного нормального распределения (например, для α=0.05, $z \approx 1.96$)*

В формуле можем влиять только на дисперсии!

Способы уменьшения дисперсии

- Повышение качества данных
- Фильтрация выбросов
- Стратификация
- CUPED

Стратификация

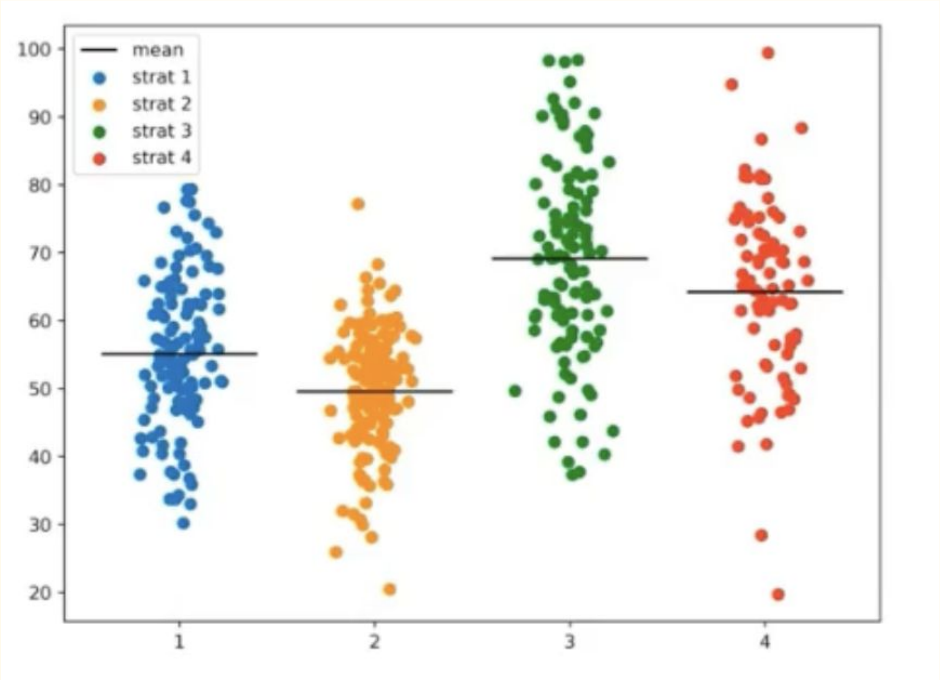
Предположим, что нам удалось найти один или несколько признаков, которые коррелируют с исследуемой бизнес метрикой Y. Такие признаки X мы будем называть **ковариатами**. Эти величины должны быть измеримы до эксперимента.

Например, это могут быть **пол, возраст или иные характеристики пользователя**. Для международных онлайн-платформ хорошим признаком будет страна проживания пользователя.

Ковариаты используются для того, чтобы разделить всю генеральную совокупность на K непересекающихся подмножеств, называемых **стратами**.

Стратификация

- ***Суть стратификации*** – поделить генеральную совокупность на страты, в которых отличается среднее
- При стратификации среднее значение не меняется, а дисперсия среднего снижается



Стратификация

- Самый лучший вариант – использовать в АБ-тесте стратифицированные группы.
- Вариант похуже – если забыли стратифицировать до АБ, стратифицируйте после АБ на постанализе:)

- Стратификация помогает не всегда. Не всегда найдутся ковариаты, которые будут снижать дисперсию.
- До АБ можно потестировать, как разбиение генеральной совокупности по определенным признакам снижает дисперсию среднего.

CUPED

CUPED (Controlled Experiments Using Pre-Experiment Data) – это статистический метод, позволяющий уменьшить дисперсию метрик в A/B-тестах за счёт использования исторических (pre-experiment) данных.

- Наблюдаемая дисперсия частично обусловлена неустранимым разбросом и частично связана с влиянием ненаблюдаемых нами факторов.
- Если есть основания полагать, что эти факторы постоянные, тогда они также влияли на исторические данные.
- Если определить связь между историческими данными и данными в эксперименте, то дисперсию можно уменьшить.

$$Y_{\text{adjusted}} = Y - \theta \cdot (X - \bar{X})$$

$$X : \textit{ковариата},$$

$$\bar{X} : \textit{среднее значение ковариаты},$$

$$\theta : \textit{коэффициент регрессии, вычисляемый как } \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

CUPED

<u>Свойства:</u>

- При оценке неизвестного эффекта оценка останется несмещённой, если изначально была таковой.
- Дисперсия не увеличивается, а чаще снижается.
- Всё, что нужно для работы метода, это найти скоррелированный ряд.
- Легко объяснять бизнесу.

5. Практика: часть 2

Спасибо
за внимание!