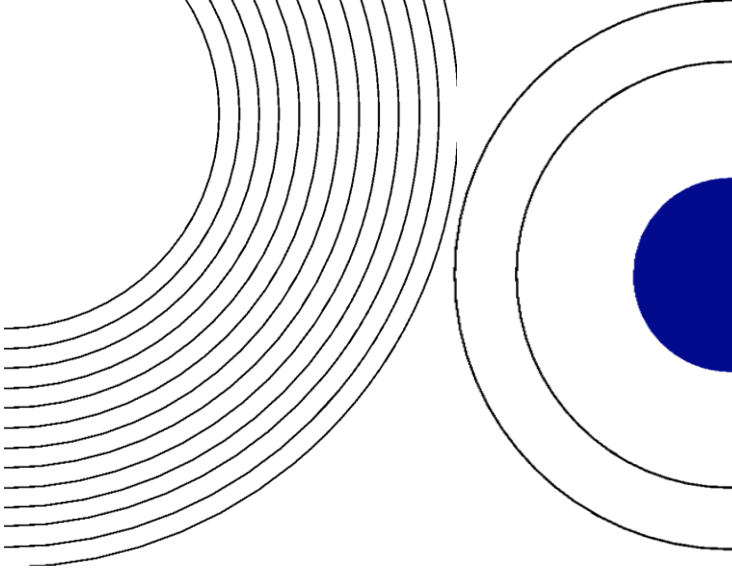
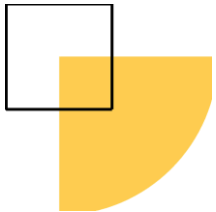
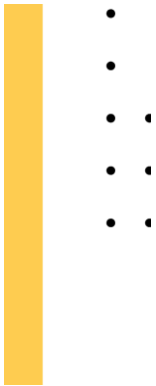
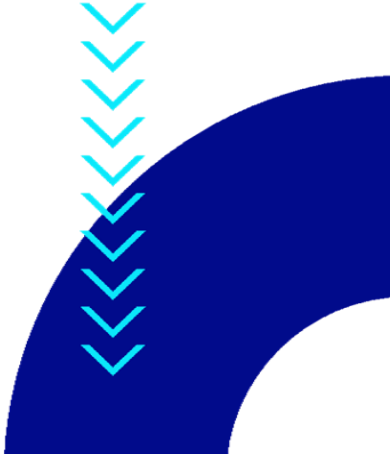
Занятие №3



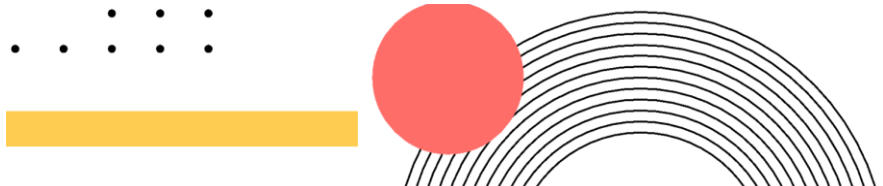




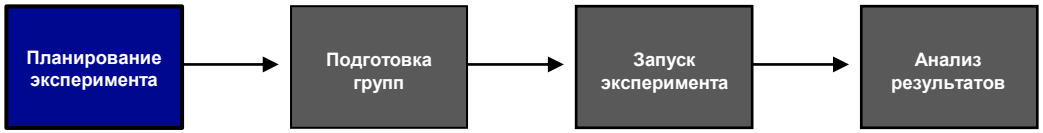


В ходе третьего занятия:

- узнаем, как определить размер групп методом имитации;
- обсудим способы снижения ошибок 1-го и 2-го рода: работа с выбросами и стратификация.

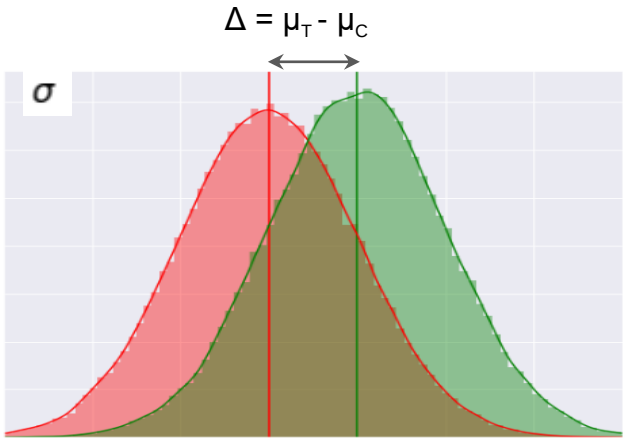


Расчет количества наблюдений: имитация Метрика — среднее



Зависимость необходимой численности групп от дисперсии данных и величины ожидаемого эффекта

1. Зафиксируем α, β
1. Зафиксируем σ, Δ
1. Фиксируем sample_size
1. Из распределений “достаем” подвыборки размером sample_size
5. Сравниваем подвыборки t-test с заданным уровнем α



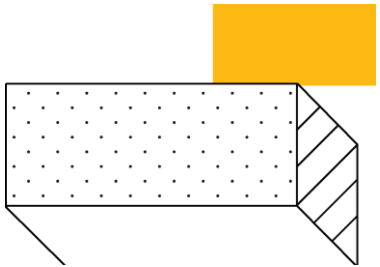
Повторяем N раз.
Считаем мощность

Повторяем для разных sample_size

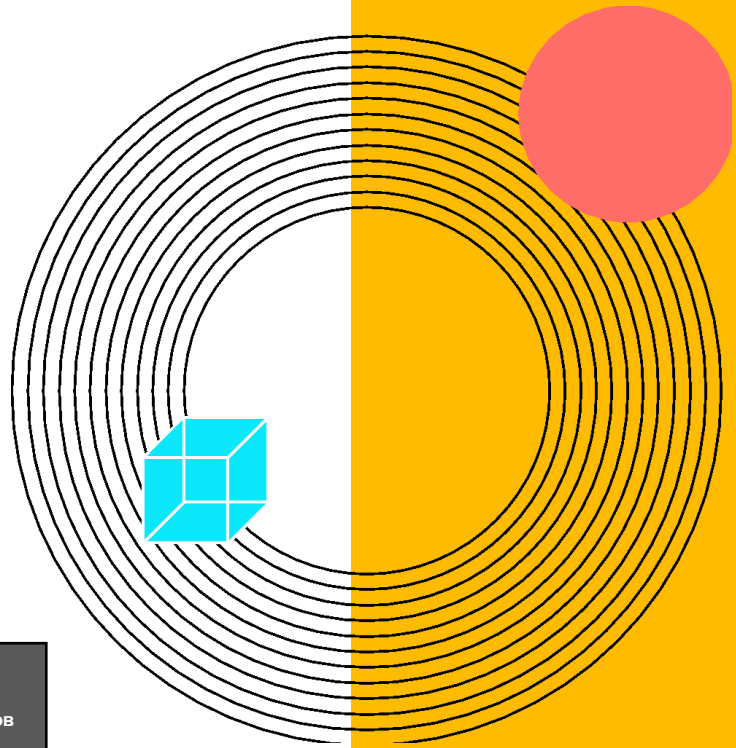
Повторяем для разных σ, Δ

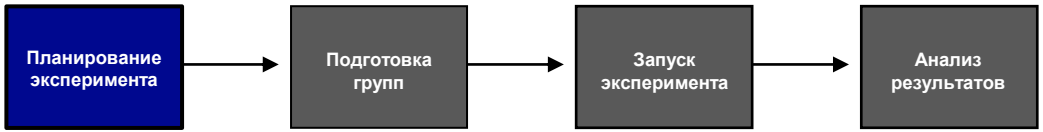
Демонстрация

Расчет размера выборки для эксперимента на нормальном распределении методом имитации



Расчет количества наблюдений: имитация Метрика — персентиль





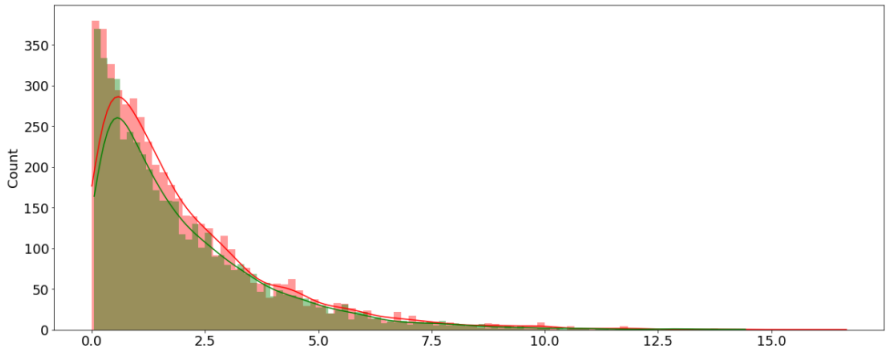
Расчет количества наблюдений на эксперименте с метрикой — 10ый персентиль

1. Зафиксируем α, β;
2. Фиксируем sample_size;
2. Фиксируем Δ;
2. Берем историческое распределение данных, формируем из него две группы размером sample_size;
4. Моделируем результаты эксперимента: к значениям одной из групп прибавляем инкремент размера Δ;
6. Сравниваем значения метрики в подвыборках через bootstrap-тест и принимаем решение на уровне alpha.

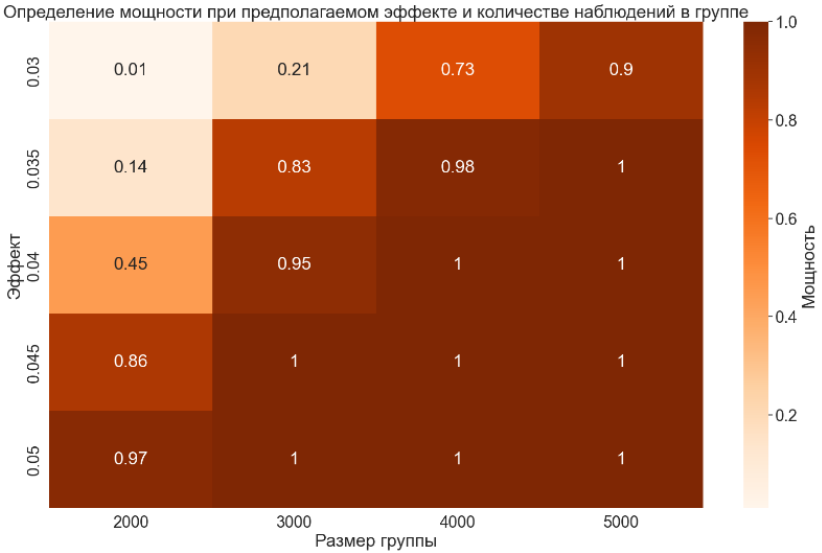
Повторяем много раз. Считаем мощность.

Повторяем для разных значений sample_size.

Повторяем для разных значений Δ.



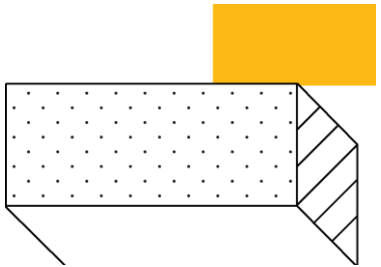
Итоги имитации

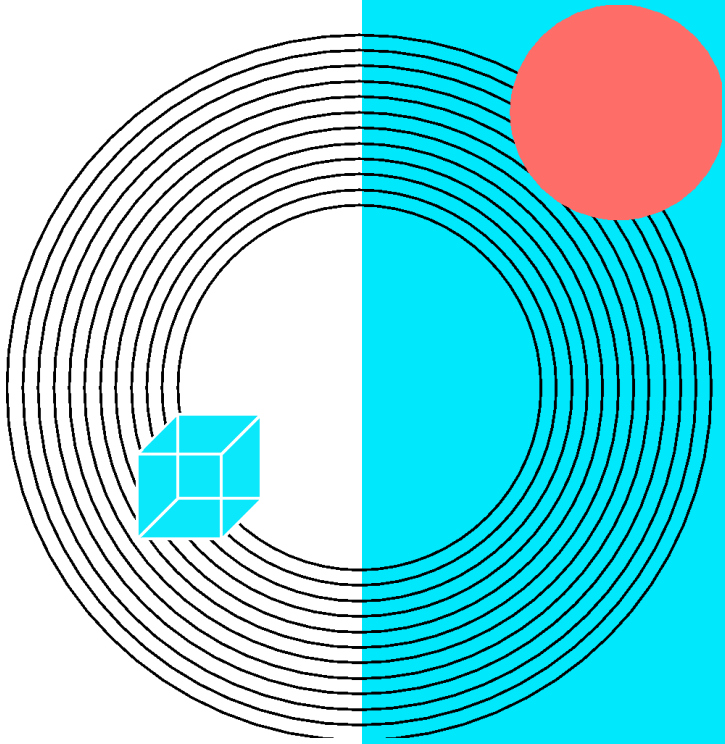


- Задав на начальном этапе уровень β (или уровень требуемой мощности), решение об окончательном варианте beta принимается после подведения итогов по имитации
- Если в эксперименте подразумевается неодинаковое соотношение размера групп, то это так же закладывается на моменте имитации
- Неодинаковый размер групп в сумме требует большее число наблюдений на эксперименте

Демонстрация

Расчет размера групп для произвольного распределения

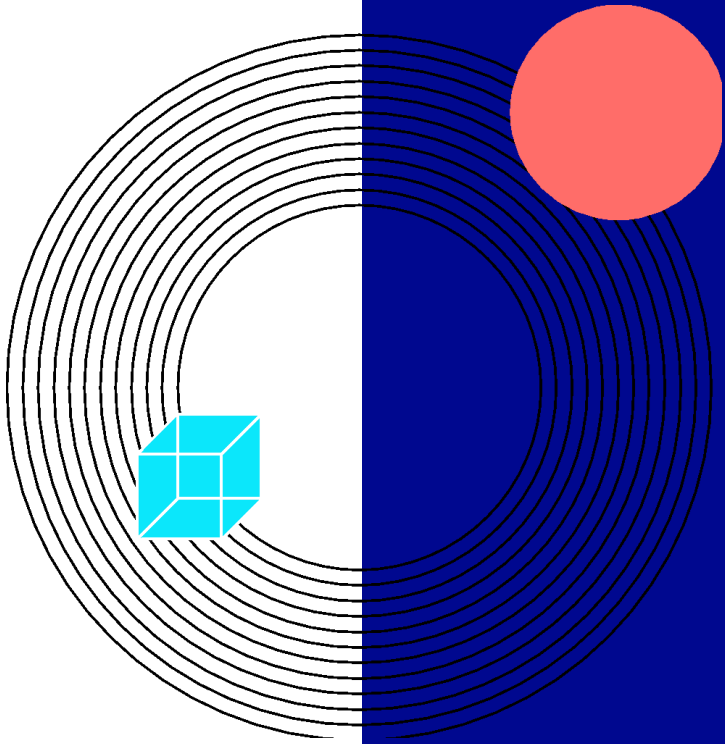




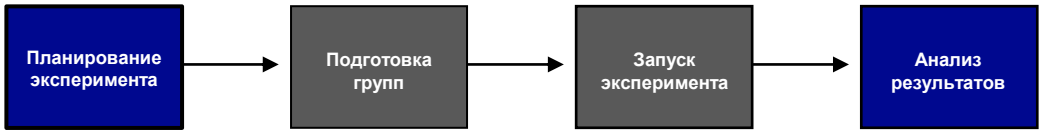
Демонстрация

Расчет размера данных для эксперимента по формуле.

Сравнение подходов к расчету параметров эксперимента.

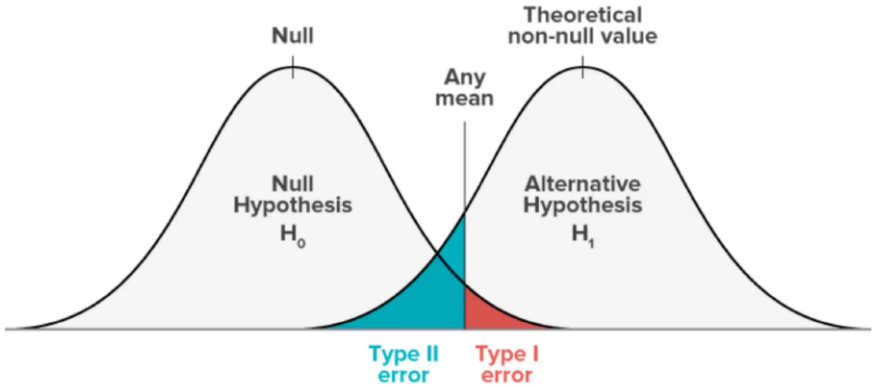


Способы уменьшения ошибок 1 и 2 рода



Вспомним определение ошибки 1-го и 2-го рода

		Правда			
		H0		H1	
Гипотеза, в сторону которой склонился тест	H0		Мы не отвергаем нулевую гипотезу.	Ошибка второго рода	Мы не отвергаем нулевую гипотезу по тесту,.
	H1	Ошибка первого рода	Мы отвергаем нулевую гипотезу согласно тесту.		Мы отвергаем нулевую гипотезу.



Ошибки 1-го и 2-го рода взаимосвязаны

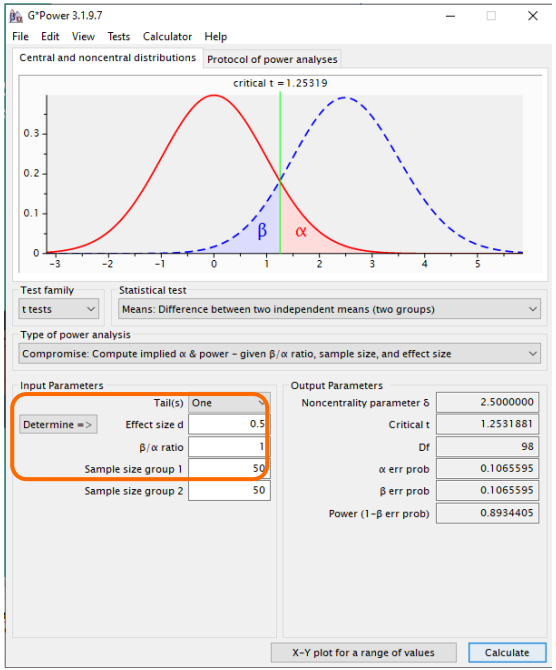


Рис. 1.: соотношение ошибок стоит 1 к 1. И мы видим, что при заданном размере групп и эффекте, это означает, что ошибка 1-го рода 0.1 и ошибка 2-го рода 0.1, а мощность почти 0.9.

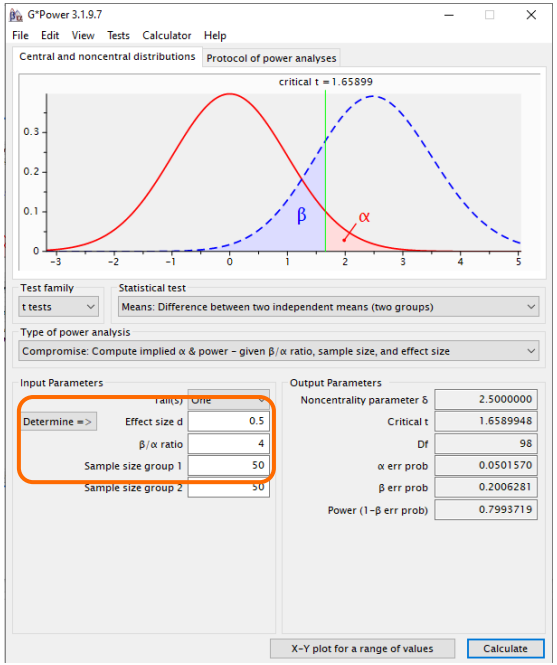
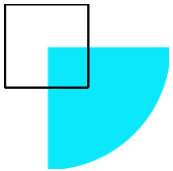


Рис. 2.: соотношение ошибок стоит на стандартном уровне 1 к 4. При заданном размере групп и эффекте, это означает, что ошибка 1-го рода 0.05 и ошибка 2-го рода 0.2, а мощность почти 0.8.

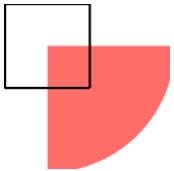
Функция Compromise калькулятор **G*Power**

Где скачать калькулятор: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

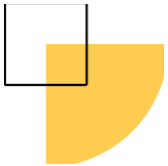
На каком уровне должны быть ошибки 1-го и 2-го рода?



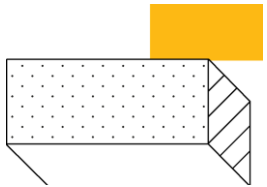
Сколько стоит переход на новое решение, которое тестируем?



Какие риски сопряжены с переходом на новое решение?

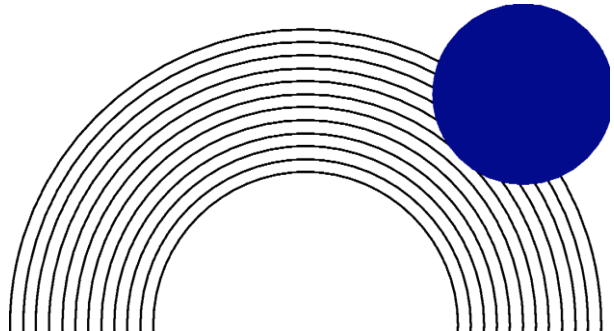


Какие риски сопряжены с тем, что мы останемся со старым решением?

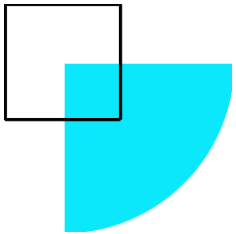


К способам уменьшения ошибок 1-го и 2-го рода относят

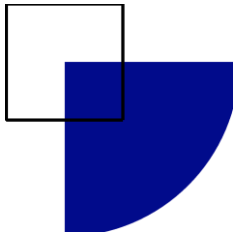
- Увеличение размера групп
- Взятие групп максимально равного размера
- Уменьшение дисперсию: убрать выбросы или привести распределение к нормальному
- Применение стратификации при формировании групп, чтобы гарантировать их гомогенность
- Уменьшение дисперсии: CUPED и другие методы



Зачем оптимизировать дизайн теста с оглядкой на ошибку 1-го и 2-го рода?

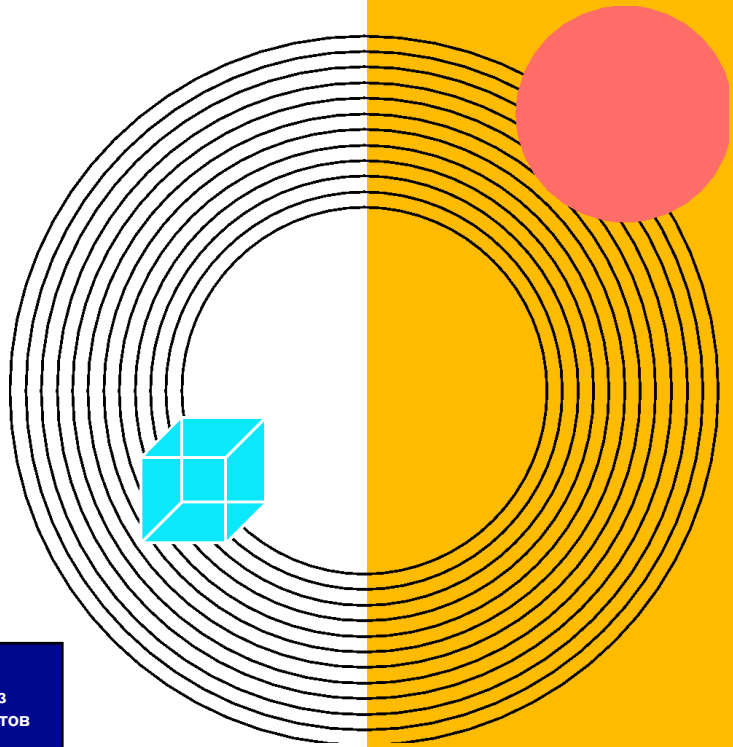


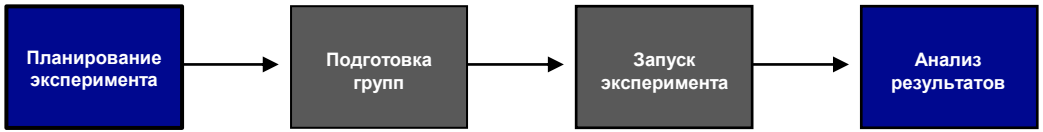
Уменьшение числа наблюдений



Уменьшение длительности теста

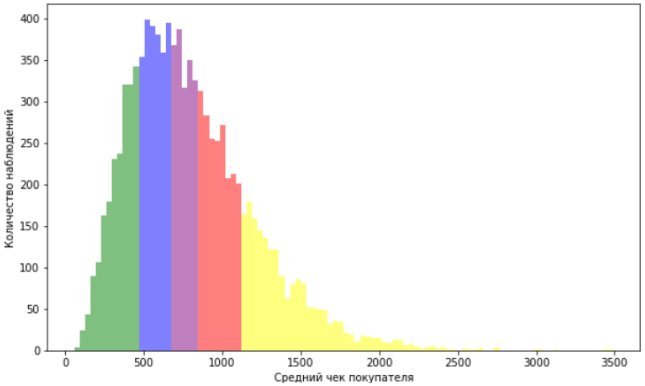
Стратификация



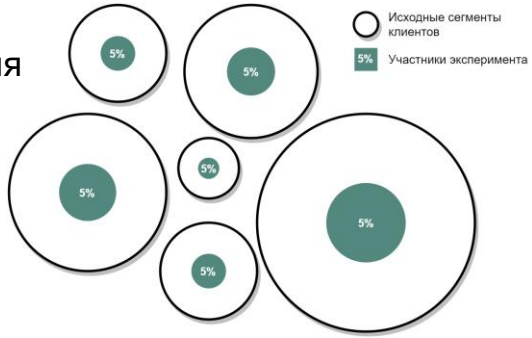


Способ 1. Стратификация

Бакетирование / сегментация (e.g. RFM)



Кластеризация



Выполняет две функции:

- Когда хотим набрать сопоставимые группы.
- Когда хотим проверить сопоставимость групп.

Плюсы: идея подхода логична, просто объясняется бизнесу и научно обоснована.

Минусы / сложности: от выбора метода и переменных, по которым осуществляется стратификация, зависит качество.

Условно есть два подхода:

- Простое пересечение переменных (если переменные непрерывные - их бакетируют);
- Кластеризация клиентов при помощи любого ML-алгоритма.

Валидация стратификации

1 Используем алгоритм классификации

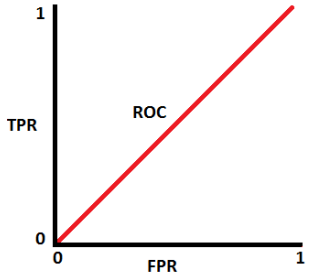
Если качество модели плохое, то сплитовалка хорошая!

Что значит “плохое качество модели”?

#объекта	набор фичей	группа
1	X1	1
2	X2	1
3	X3	0
4	X4	0
5	X5	1

(А) ROC AUC = 0.5

(Б) Логистическая регрессия: коэффициенты одинаково значимы (или одинаково незначимы).



Алгоритм :

- В качестве фичей берем переменные, которые кажутся нам значимыми с бизнес точки зрения;
- Строим логистическую регрессию, в которой целевое значение - 0 или 1 в зависимости от группы А и Б;
- Если хотя бы одна из фичей оказывается значимой для предсказания целевого события, то есть модель различает группы, то есть проблемы с гомогенностью.

Валидация стратификации

2 Chi-squared test

Важно: не дает ответа на вопрос: по какой из страт отличаются группы?

Отвечает на вопрос: Есть ли отличия?

Для каждой группы берем:

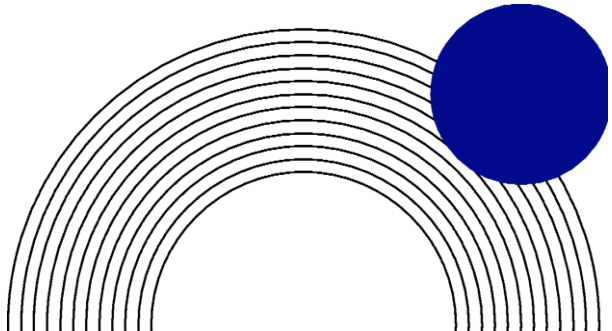
- значение, которое мы видим.
- ожидаемое значение (то есть значение другой группы);
- вычитаем одно из другого и возводим в квадрат;
- делим на ожидаемое значение;
- суммируем.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2 = chi squared

O_i = observed value

E_i = expected value



Валидация стратификации

3 PSI - population stability index

$$PSI = \sum \left((Actual\% - Expected\%) \times ln(\frac{Actual\%}{Expected\%}) \right)$$

Breakpoint Value	Bucket	Initial Count	New Count	Initial Percent	New Percent	PSI
-2.330642	1	1	0	0.01	0.001000	0.020723
-1.801596	2	1	3	0.01	0.025000	0.013744
-1.272550	3	4	6	0.04	0.050000	0.002231
-0.743504	4	8	15	0.08	0.125000	0.020083
-0.214458	5	27	18	0.27	0.150000	0.070534
0.314588	6	22	23	0.22	0.191667	0.003906
0.843633	7	16	26	0.16	0.216667	0.017181
1.372679	8	12	14	0.12	0.116667	0.000094
1.901725	9	6	9	0.06	0.075000	0.003347
2.430771	10	3	3	0.03	0.025000	0.000912

Интерпретация:

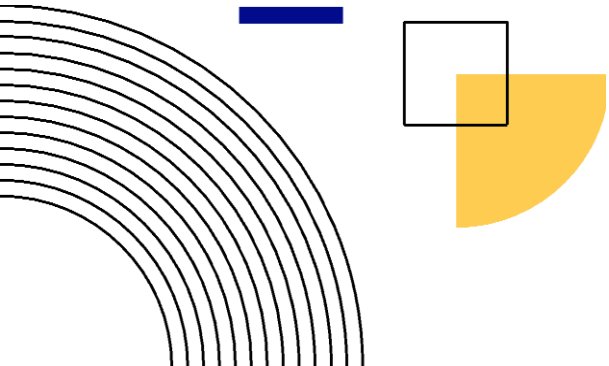
- PSI < 0.1 - значимых различий нет;
- PSI < 0.2 - есть незначительные различия;
- PSI ≥ 0.2 - есть значительные различия.

<<< Пример проверки качества модели.

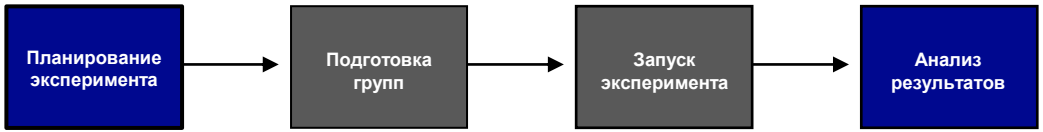
Взято из: Population stability index. URL: https://mwburke.github.io/data%20science/2018/04/29/population-stability-index.html

Демонстрация

Стратификация



Работа с выбросами



Способ 2. Работаем с выбросами

Убираем выбросы или подменяем их на максимально “разумное” значение

Когда используем:

- Выбросы объясняются техническими ошибками (грязные данные);
- Выбросы — это клиенты из сегмента, с которым мы не работаем в тесте.

Сколько убирать:

- минимально (идеально не более 1 проц. пункта).

В чем минус:

- иногда сложно договориться с бизнесом о том, что считать “выбросом”.

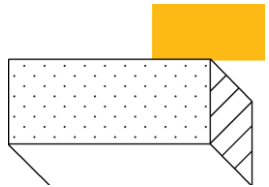
Меняем форму распределения (например, логарифм)

Когда используем:

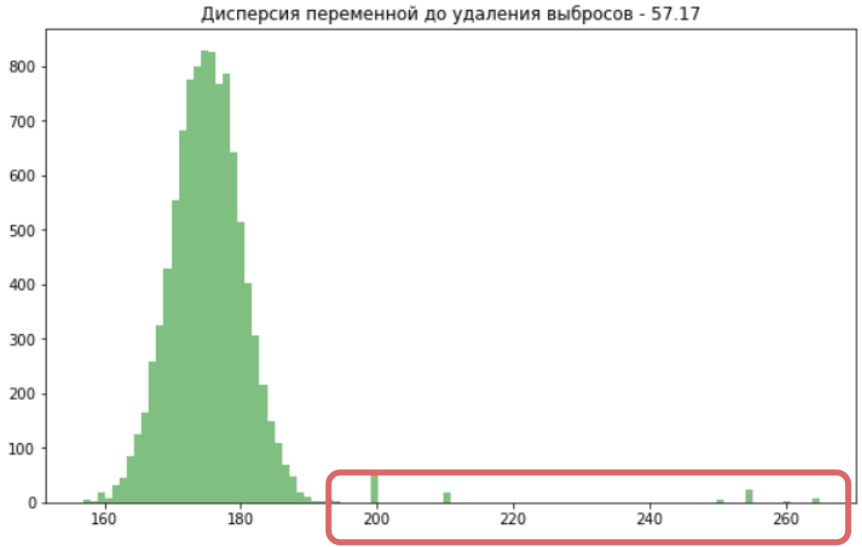
- Когда уверены, что трансформация метрики не меняет значимо интерпретацию теста для бизнеса.

В чем минус:

- Тест почти всегда становится менее прозрачным, а значит, к нему меньше доверия со стороны бизнеса.



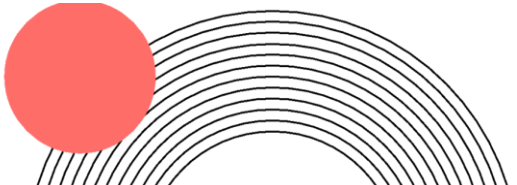
Способ 2. Работаем с выбросами



Убрали наблюдения с z-score > 3.
Дисперсия уменьшилась более, чем в 2 раза.

Выводы по третьему занятию

- Для расчета количества наблюдений для эксперимента можно пользоваться как формульным, так и имитационным подходом, причем последний более предпочтительный.
- Ошибки первого и второго рода взаимосвязаны.
- Есть несколько способов, которые позволяют снизить необходимый размер групп и/или длительность теста (это же подразумевается, если мы говорим о снижении ошибке второго рода). По сути они все направлены на снижение дисперсии.
- Успешность A/B теста зависит от подготовки эксперимента в целом и конкретно от того, на сколько удалось набрать одинаковые группы. Следует уделять внимание их гомогенности.
- Для проверки гомогенности существует несколько способов. Удобно настроить один и пользоваться им для всех тестов на одном продукте.



Литература

- Доверительное A/B-тестирование. Практическое руководство по контролируемым экспериментам (2021).
- Increasing experimental power with variance reduction at the BBC;
- Five ways to reduce variance in A/B testing;
- Увеличение чувствительности А/Б-тестов с помощью Cuped. Доклад в Яндексе;
- How Booking.com increases the power of online experiments with CUPED;
- Online Experiments Tricks — Variance Reduction.

ВОПРОСЫ