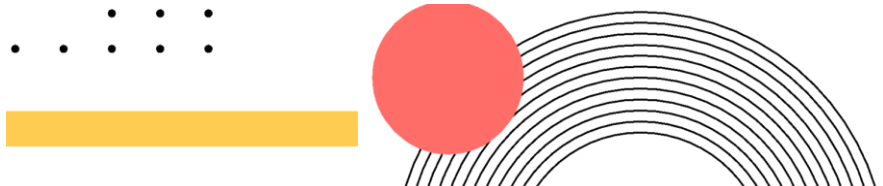
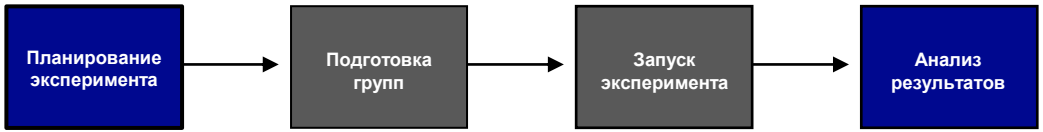
Занятие №2

В ходе второго занятия:

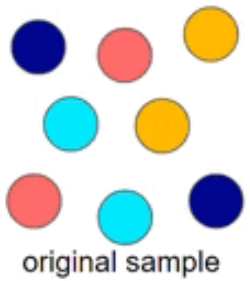
- проговорим, как используется Bootstrap;
- разберемся с ошибками I и II рода, научимся находить компромисс между ними;
- узнаем про минимальный детектируемый эффект и его интерпретируемость;
- научимся определять размер групп для эксперимента;
- узнаем, как определить размер групп, если их численности разные;
- поймем, как находить стандартное отклонение метрики для рачета размера групп.



Bootstrap


Планирование эксперимента
Подготовка групп
Запуск эксперимента
Анализ результатов

Bootstrap



bootstrap sample 1

bootstrap sample 2

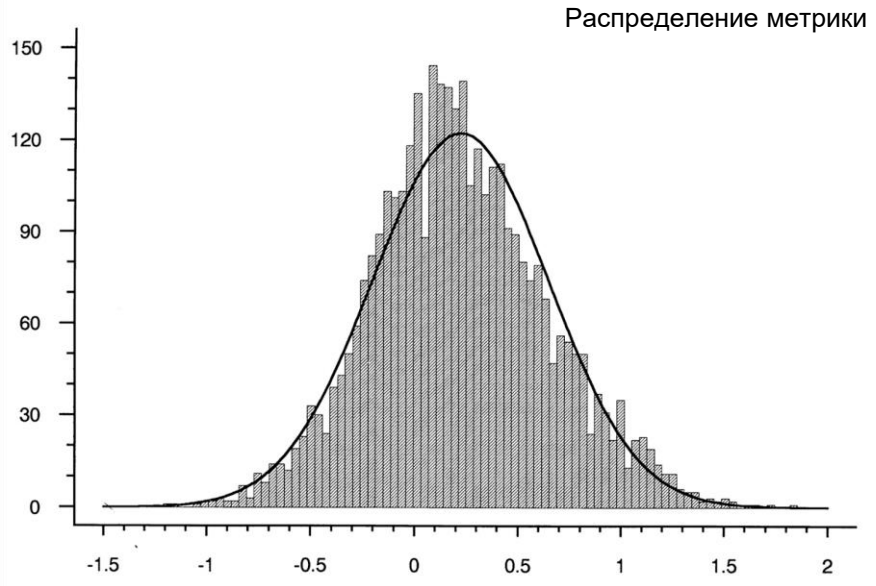
bootstrap sample 3

Ресэмплинг: имитация повторного эксперимента (имитация повторной выборки из генеральной совокупности)

Описать каждое из полученных распределений одним числом (например, средним или медианой)

Распределение значений метрики для bootstrap-сэмплов

- На каждом сэмпле вычислим искомую метрику
- Повторяем ресэмплинг - вычисляем значение метрики - строим гистограмму
- Большинство полученных нами значений метрики (при многократной имитации повторного эксперимента) будет сосредоточено вблизи истинного значения метрики
- То есть истинному значению метрики будет соответствовать **мода** полученного нами распределения значений метрики



Bootstrap-тест

Ресэмплинг из группы A и из группы B

Описать каждое из полученных распределений одним числом (например, средним или медианой)

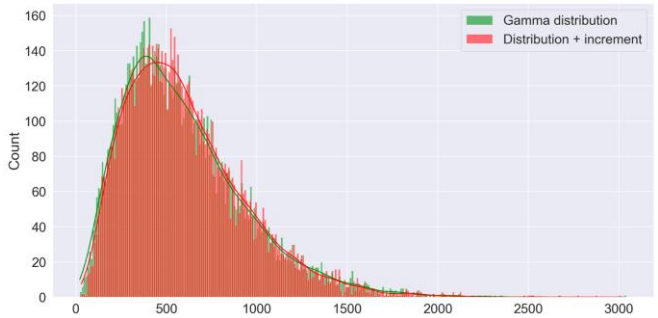
Вычислить разность полученных чисел

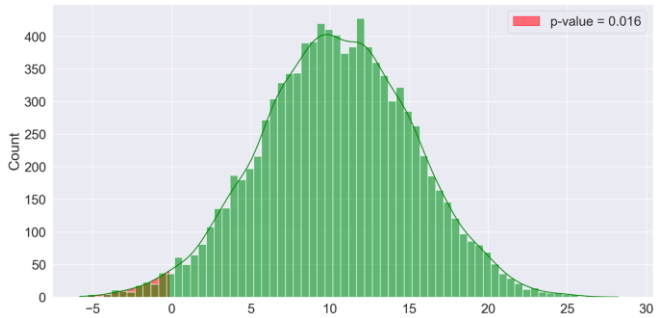
Сформировать распределение таких разностей

Вычислить статистическую значимость нуля в распределении таких разностей

Если исходные распределения ничем не отличаются, то количество положительных разностей будет, примерно, равно количеству отрицательных разностей - и статистическая значимость нуля будет около 100%

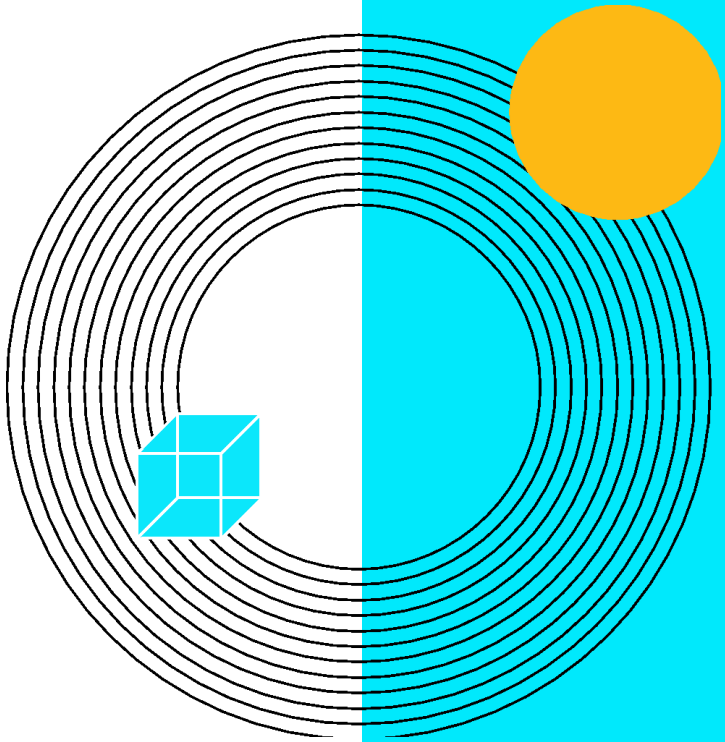
Если же исходные распределения различаются, то статистическая значимость нуля будет ниже и, возможно, существенно ниже 100%



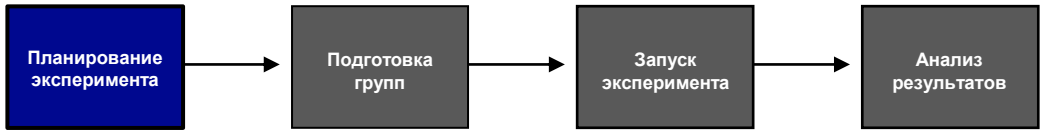


Демонстрация

Расчет p-value через bootstrap-тест



Ошибки 1 и 2 рода



Ошибка I рода

Ошибка первого рода α— отвергаем нулевую гипотезу, хотя она в действительности верна.

Ошибка первого рода α— замечаем различие там, где его нет.

Продуктовая постановка: выкатываем фичу в продакшн, хотя фича не увеличивает требуемую метрику.

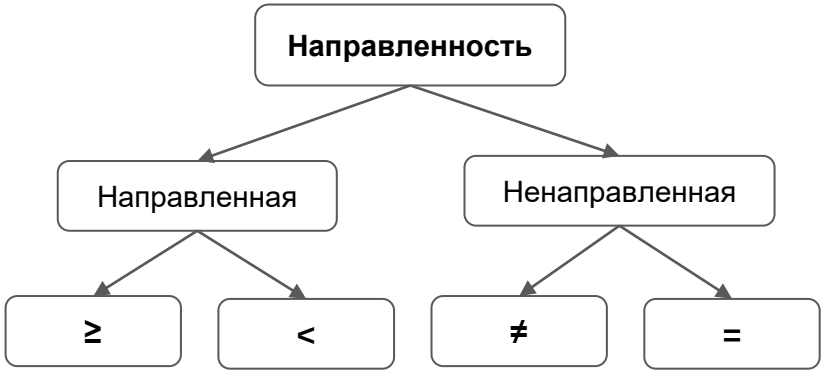

```
n_samples = 10_000
sample_size = 10_000

alpha = 0.05
real_alpha = 0

for _ in range(n_samples):
    a = np.random.normal(0, 3, sample_size)
    b = np.random.normal(0, 3, sample_size)

    test_res = ttest_ind(a, b, equal_var=False)
    if test_res[1] < 0.05:
        real_alpha += 1

real_alpha /= n_samples

print(f'Theoretical alpha: {alpha}')
print(f'Real alpha: {real_alpha}')
```

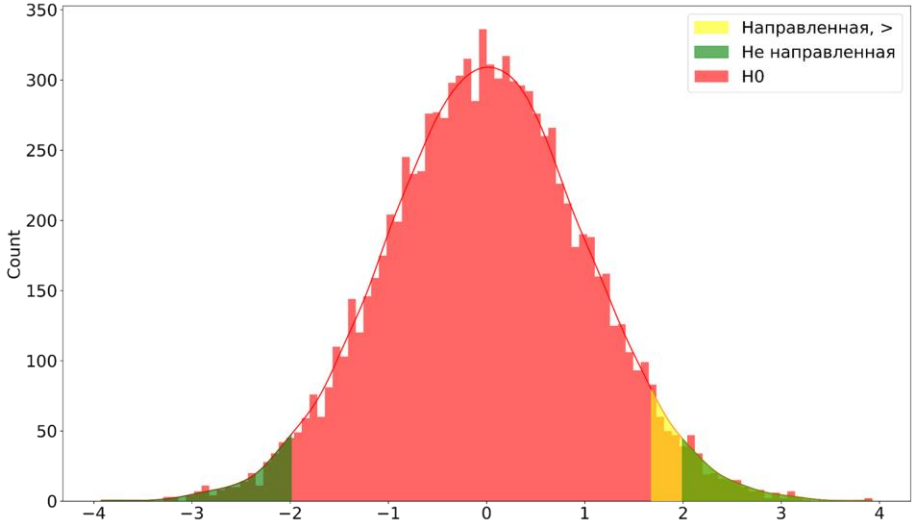
```
Theoretical alpha: 0.05
Real alpha: 0.049
```

Если между распределениями нет отличий, то мы будем получать различия в **α%** случаев.

Направленность гипотезы

Направление гипотезы позволяет сделать вывод о том, присутствуют ли отличия в целом или можно явно сказать об их направленности (отличия “больше” или “меньше”).

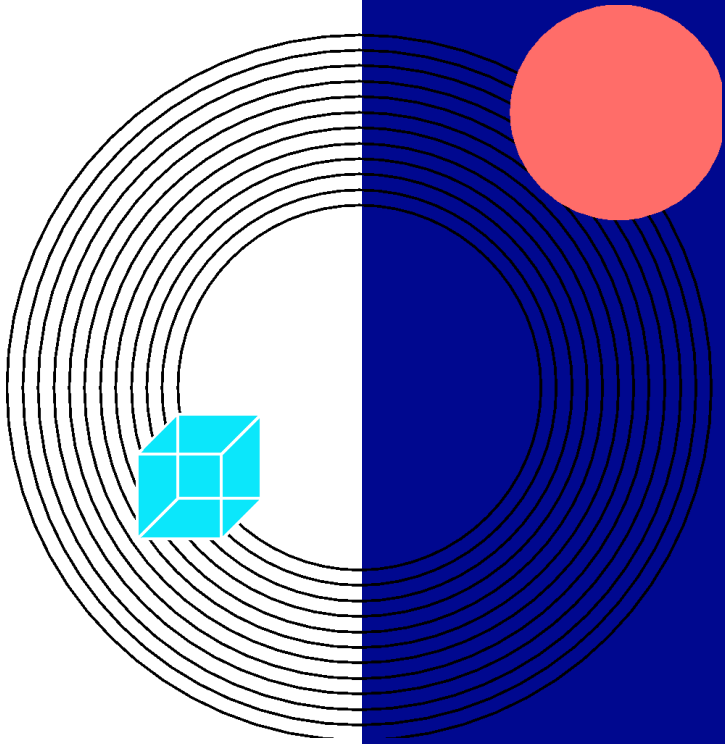




Вывод: при направленной гипотезе увеличивается вероятность отклонения нулевой гипотезы.

Демонстрация

Ошибка I рода



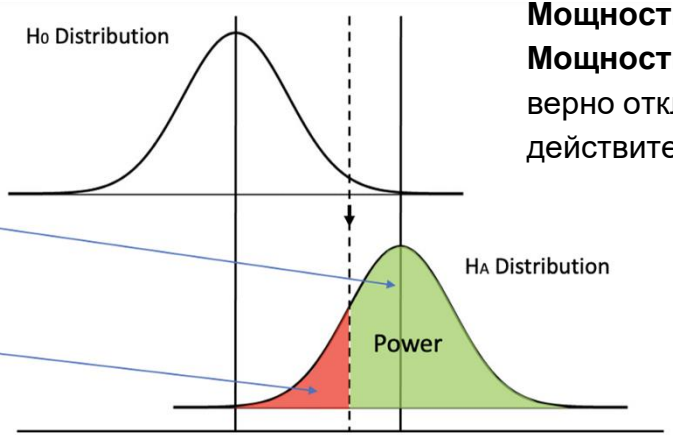
Ошибка II рода

Ошибка второго рода β — не отвергаем нулевую гипотезу, хотя она в действительности не верна.

Ошибка второго рода β — не замечаем различия там, где оно есть.

Продуктовая постановка: разработанная фича оказалась полезной, но в продакшн мы ее не выкатили.

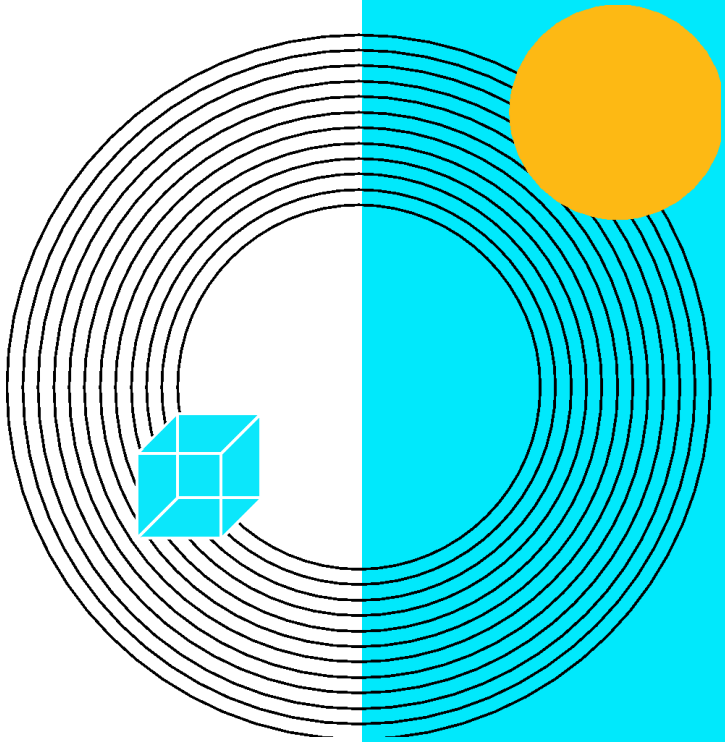
	H_0 is True	H_A is True
Reject H_0	Type I Error False Positive α (alpha)	Correct True Positive 1-β (power)
Fail to reject H_0	Correct True Negative 1-α	Type II Error False Nagative β (beta)



Мощность теста = 1 - ошибка второго рода
Мощность теста — вероятность того, что тест верно отклонит нулевую гипотезу, когда в действительности верна альтернативная.

Демонстрация

Ошибки I и II рода



Ошибки I и II рода - резюме

Null hypothesis is ...	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1 - β
Not rejected	Correct decision True negative Probability = 1 - α	Type II error False negative Probability = β

Trade-off

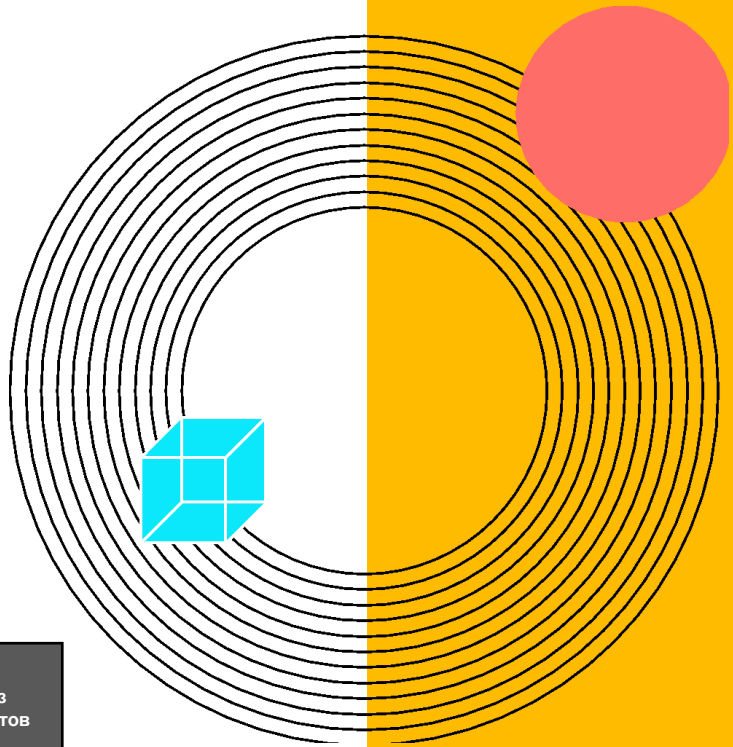
- Если поддержка фичи стоит дорого, то следует обращать **большее внимание на ошибку I рода**.
- Если есть боязнь пропустить инновацию, то следует обращать **большее внимание на ошибку II рода**.

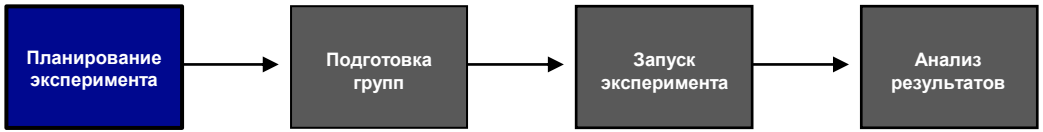
Классические ошибки:

- I рода — 5%
- II рода — 20%

Type I & Type II Errors | Differences, Examples, Visualizations

Доверительный интервал





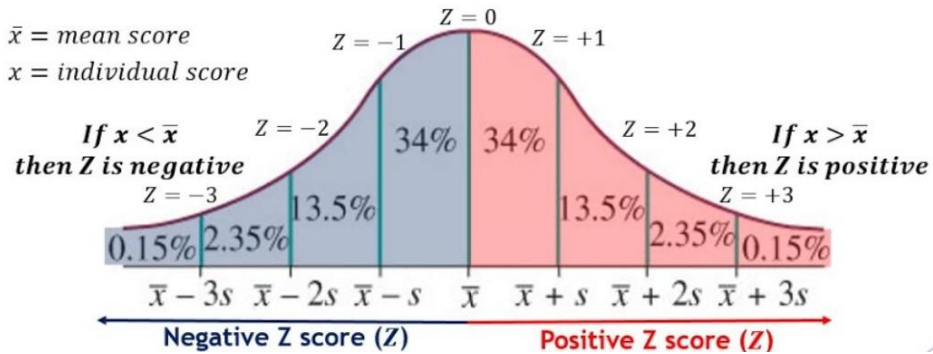
Z-score

Z-score — количество стандартных отклонений, на которое значение из ряда наблюдений выше или ниже среднего значения ряда:

$$z = \frac{x - \mu}{\sigma}$$

Для стандартного нормального распределения *N(0, 1)* **z-score — квантиль** распределения, например:

- $Z_{0.950} = Z_{1 - 0.050} = 1.64$
- $Z_{0.975} = Z_{1 - 0.025} = 1.96$



Стандартная ошибка

Стандартная ошибка статистики — величина, характеризующая стандартное отклонение выборочного среднего:

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Формула стандартной ошибки следует из центральной предельной теоремы (ЦПТ).

Пример: выборка (n = 10 000, s = 15, x_{avg} = 10). Оценить стандартную ошибку.
SE_x = 15 / 100 = 0.15

Доверительный интервал и margin of error

Margin of error (MOE) — ширина доверительного интервала.
При прочих равных при меньшем размере выборки мы получим более широкий доверительный интервал.

Дов. интервал для среднего

$$CI_{1-\frac{\alpha}{2}} = \overline{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

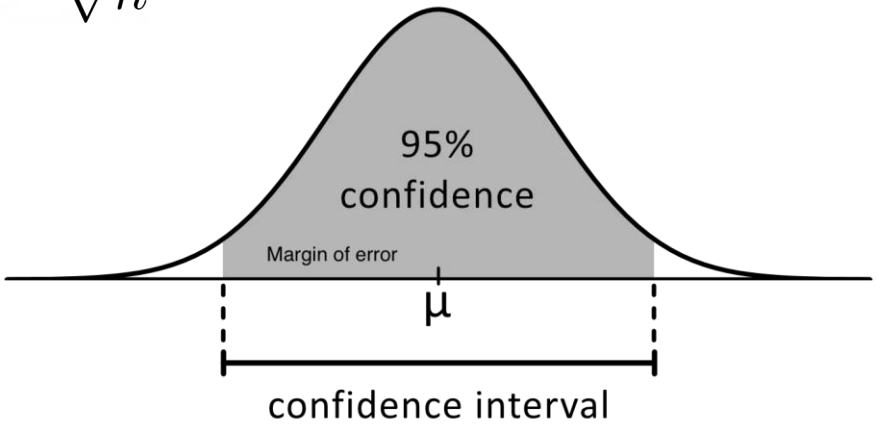
Margin of error

$$E = z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

стандартная ошибка

количество
стандартных ошибок

Power and Sample Size Determination

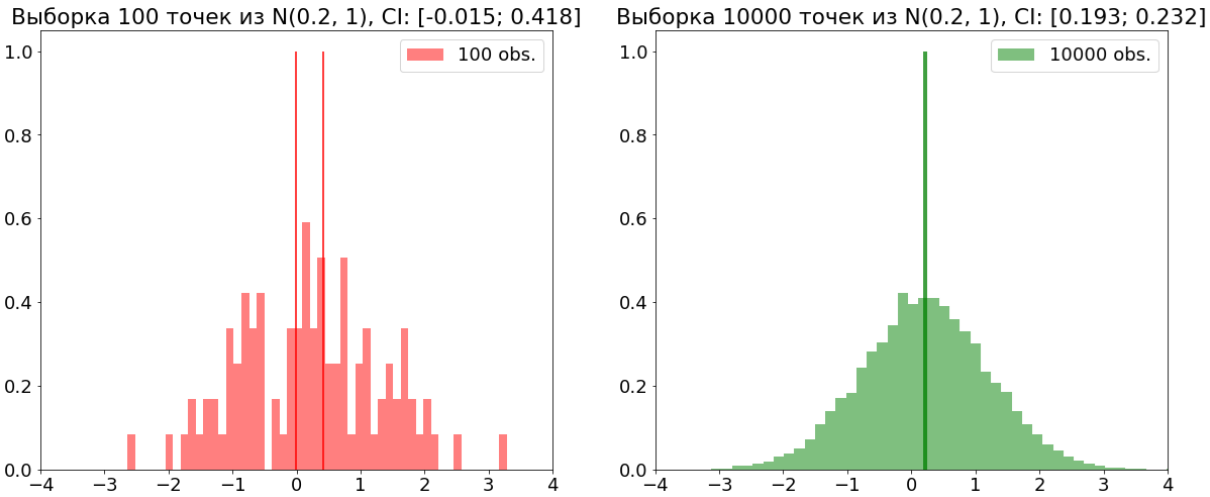


Доверительный интервал и размер выборки

Пример: имеются две выборки (100 и 10 000 точек) из нормального распределения *N(0.2, 1)*.
По ним хотим проверить гипотезу о равенстве среднего генеральной совокупности нулю:

H0: μ = 0

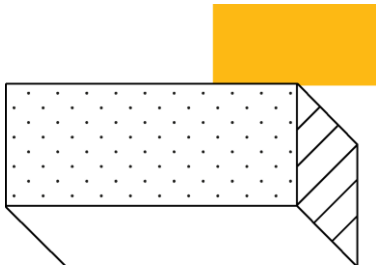
H1: μ ≠ 0



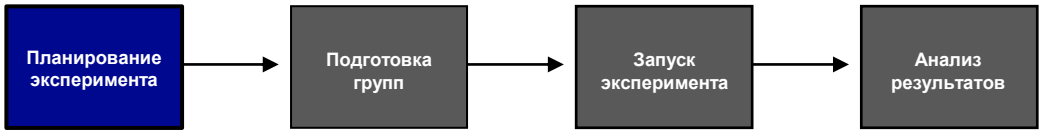
При n_1 = 100 нулевая гипотеза не отвергается, при n_2 = 10 000 — отвергается.

Демонстрация

Доверительный интервал и количество наблюдений

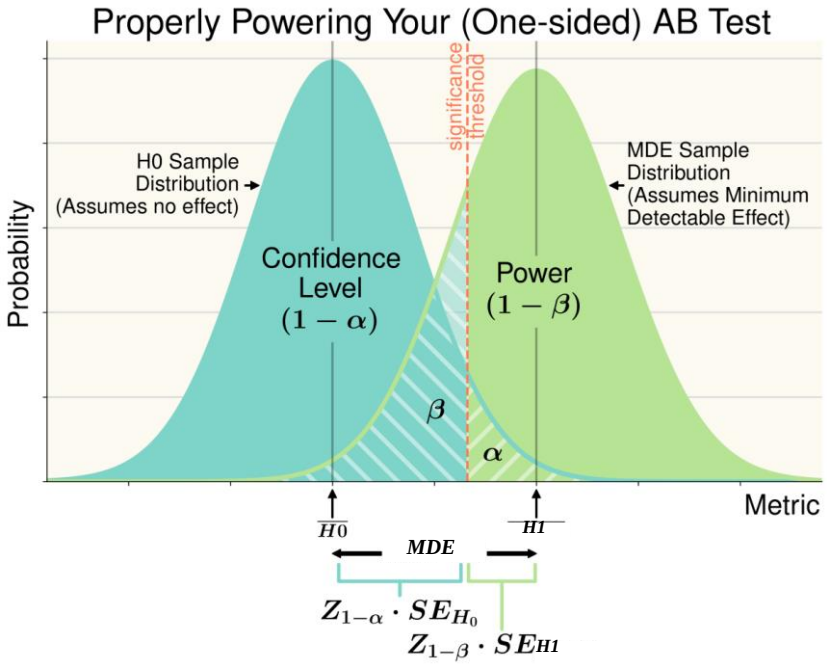


MDE — Minimum Detectable Effect



MDE — Минимальный детектируемый эффект

$$E = z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$



MDE — минимальный эффект, который готовы увидеть в эксперименте.

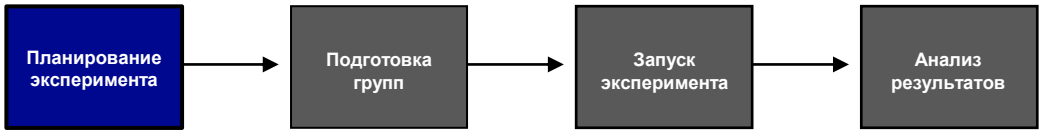
$$MDE = Z_{1-\alpha} * SE_{H0} + Z_{1-\beta} * SE_{H1}$$

На MDE влияет:

- направленность гипотезы
- **α** — ошибка первого рода
- **β** — ошибка второго рода
- σ_A, σ_B — стандартные отклонения метрик
- n_A, n_B — размеры групп

Calculating Sample Sizes for A/B Tests [article]

Расчет количества наблюдений: формула



Расчет размера выборки через формулу

Группы одинакового размера

$$n_i = 2 * \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{(\mu_T - \mu_C)^2}$$

- σ — ст. отклонение метрики
- μ_T — среднее тестовой группы
- μ_C — среднее контрольной группы

Группы разного размера

$$n * P * (1 - P) = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{(\mu_T - \mu_C)^2}$$

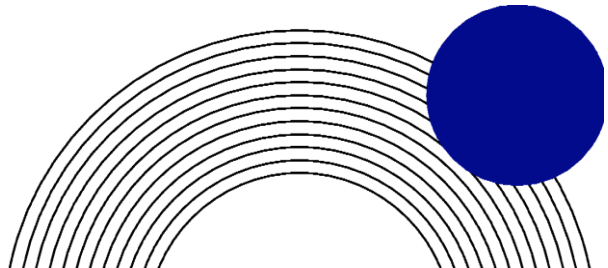
- $n = n_T + n_C$
- P — доля тестовой группы (0 < P < 1)

Пример

α = 0.05, β = 0.2, σ = 300, $\mu_T - \mu_C = 12$.

P = 0.5: $n_1 = 9800$, $n_2 = 9800$, $n_1 + n_2 = 19\,600$

P = 0.2: $n_1 = 12\,250$, $n_2 = 49\,000$, $n_1 + n_2 = 61\,250$



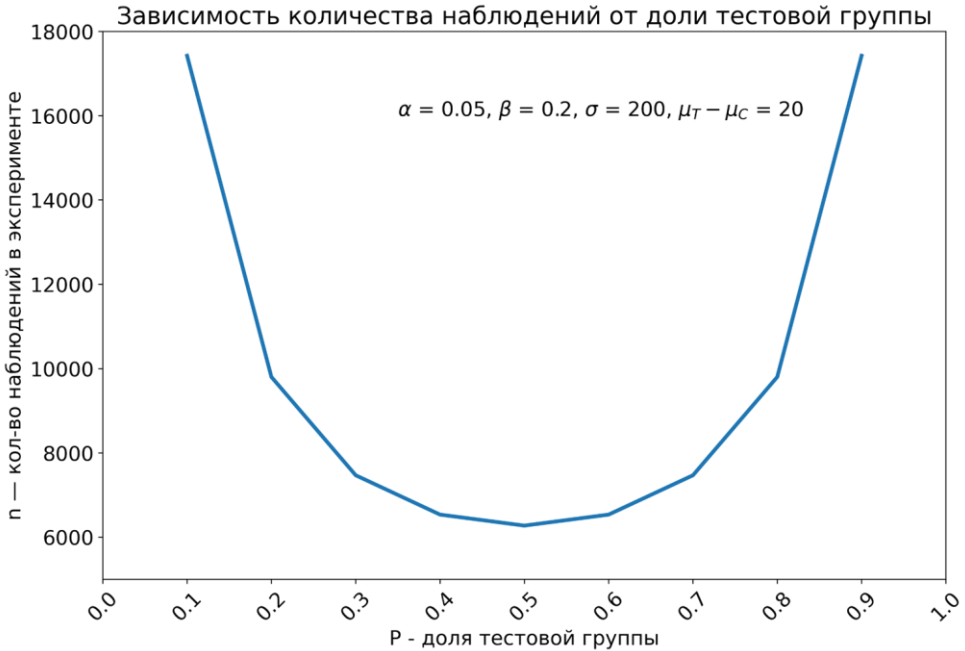
Соотношение групп эксперимента

Группы разного размера

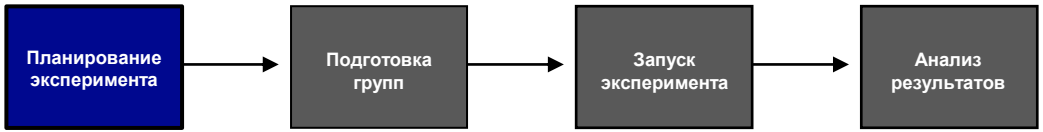
$$n * P * (1 - P) = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{(\mu_T - \mu_C)^2}$$

- $n = n_T + n_C$
- P — доля тестовой группы ($0 < P < 1$)

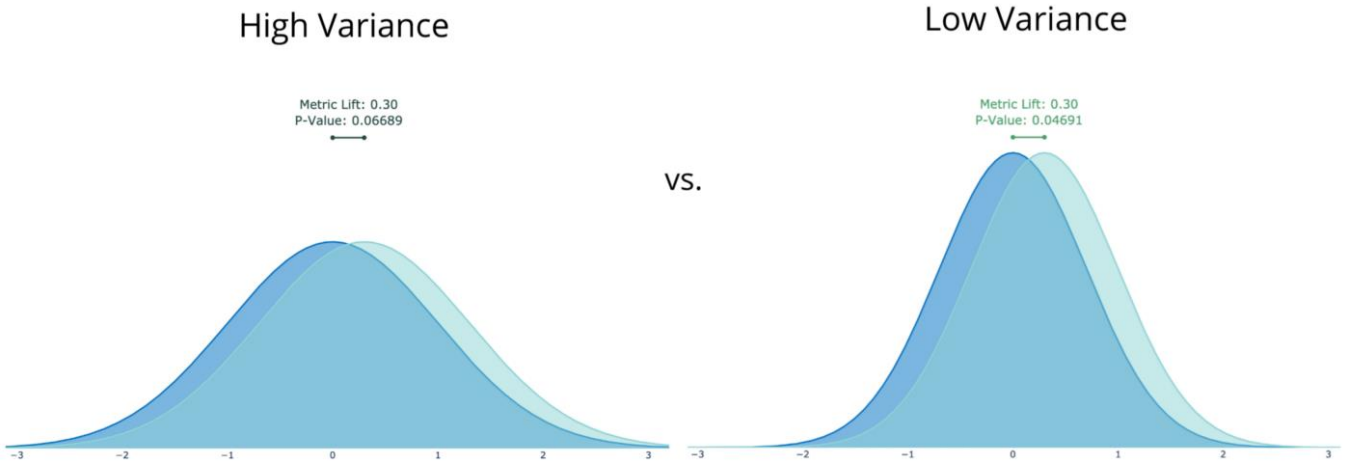
Вывод: чем более сбалансированы группы, тем меньше наблюдений требуется для эксперимента.



Стандартное отклонение и количество наблюдений



Стандартное отклонение метрики



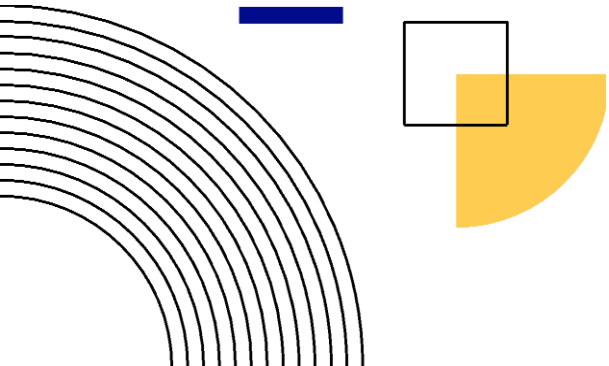
Вопрос: где взять стандартное отклонение метрики на эксперименте?

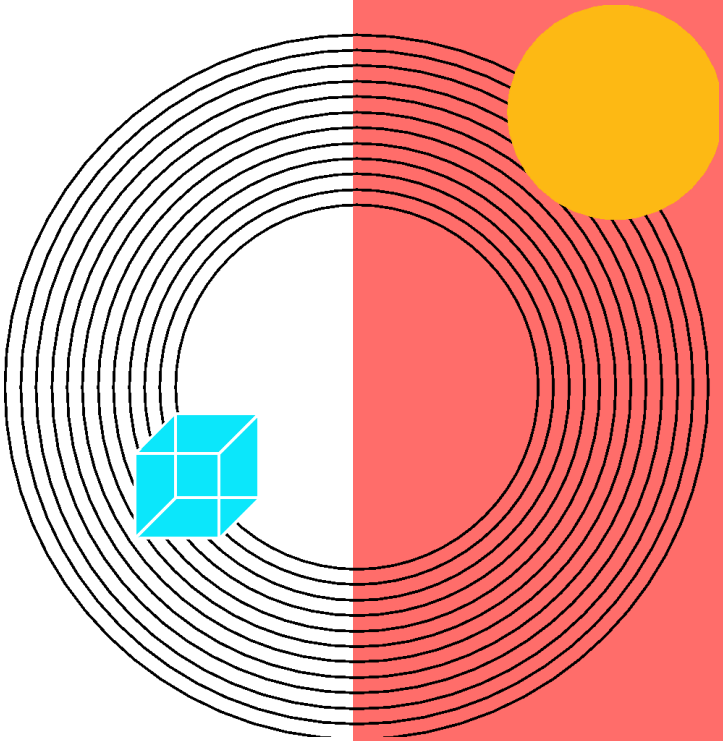
Возможные варианты:

1. Похожие эксперименты с такой же метрикой
2. Исторические данные по метрике

Демонстрация

Стандартное отклонение метрики



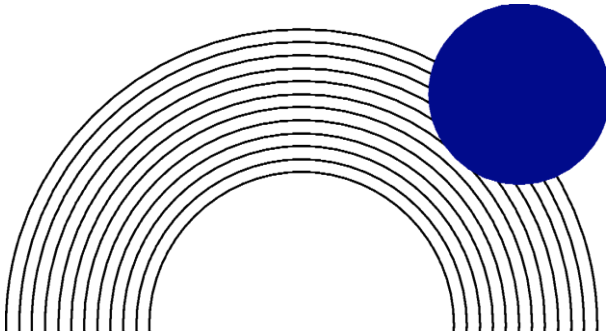


Выводы по второму занятию

- Для расчета количества наблюдений для эксперимента требуется знать следующие данные: ошибки I и II рода, стандартное отклонение тестируемой метрики, ожидаемый эффект.
- Величины ошибок I и II рода выбираются в зависимости от задачи.
- Стандартное отклонение метрики во время эксперимента может быть приближено историческими данными по аналогичной метрике.
- Большее число наблюдений в общем случае помогает увидеть более низкий ожидаемый эффект.

Дополнительная литература

- PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and QuasiExperimental Design Studies [paper]
- The Core Analytics of Randomized Experiments for Social Research [paper]



ВОПРОСЫ