



Основы А/В-тестирования

Теория и практика

Чиркина Дарья,

Аналитик, Яндекс

МГТУ им. Н.Э.Баумана,

TU Wien

Цитата-напутствие

Хороший эксперимент со статистически-значимой отвергнутой H_1 – это не проведённый эксперимент.

План лекции

1. [Практика] Business-value of A/B-test;
2. [Практика] Design A/B;
3. [Практика] Запуск эксперимента;
4. [Теория] Оценка полученных результатов: теоретические основы

1. Business-value of A/B-test

1. Этап пред-подготовки и предварительного анализа:

Выявление бизнес-целей проводимого эксперимента;

Формулировка цели. Объём аудитории.

2. Формулировка гипотез. Например, гипотеза: замена блока с иконкой сервиса «Журнал» на сниппеты (блоки с фото и кратким содержимым) статей принесёт увеличение переходов в раздел «Журнала»;

3. Оценка финансовых рисков и стоимость проведения эксперимента – преобладает ли возможный профит над фактическими затратами;

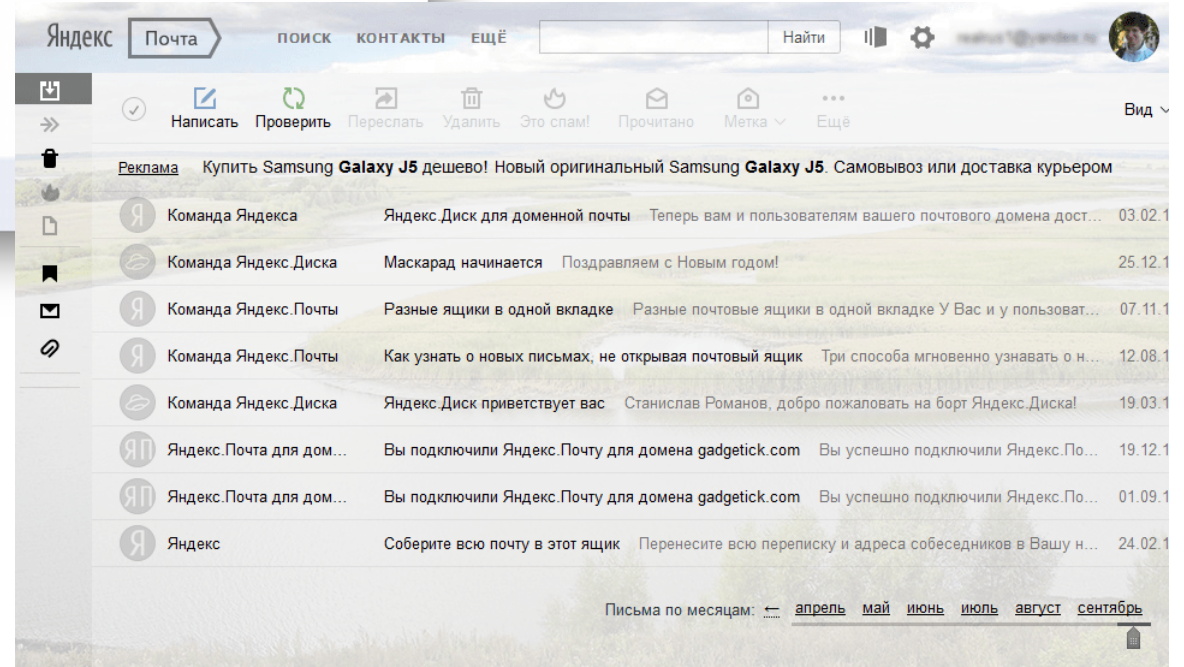
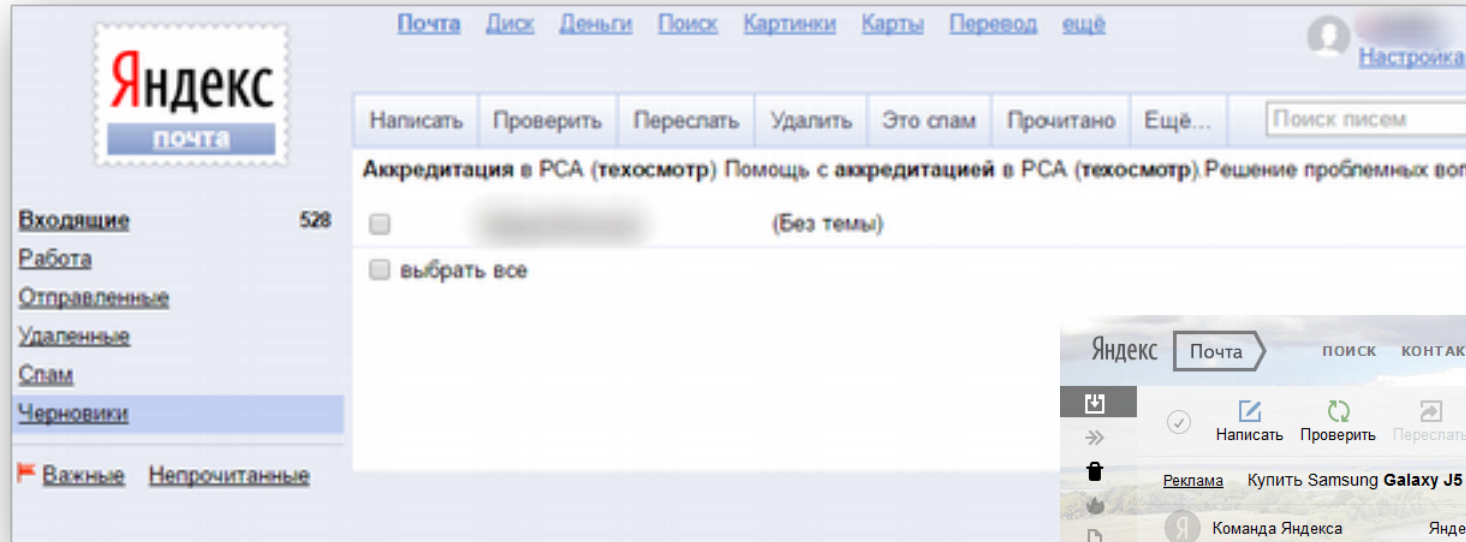
4. Исследование аудитории.

5. Пробуем оценить *стоимость* каждой гипотезы.

1. Business-value of A/B-test

- 1. Этап пред-подготовки и предварительного анализа:* выявление бизнес-целей проводимого эксперимента; Какую бизнес-цель преследуем? Формулировка цели.
Объём аудитории, которую задействует наше изменение, какой profit мы можем получить на выходе и стоит ли он затрат на проведение эксперимента и внесение изменения;
Основная идея: иногда продуктового менеджера необходимо на цифрах убедить, что планируемое изменение - дорогостоящее и невыгодное, а также его окупаемость сомнительна;
- 2. Формулировка гипотез.* Например, гипотеза: замена блока с иконкой сервиса «Журнал» на сниппеты (блоки с фото и кратким содержанием) статей принесёт увеличение переходов в раздел «Журнала»;
- 3. Оценка финансовых рисков и стоимость* проведения эксперимента – преобладает ли возможный профит над фактическими затратами;
- 4. Исследование аудитории,* которую может задействовать эксперимент (*потребности* и ёмкость рынка) – до начала проектирования эксперимента;
- 5. Пробуем оценить стоимость* каждой гипотезы – в зависимости от ожидаемого профита строим приоритеты – в каком порядке гипотезы проверять (сначала – самые профитные; оценка – по опыту: например, мы можем привлечь +10% звонков или +10% покупок и это принесёт 2 000 у.е. выручки)?

Пример: redesign почты



2. Design A/B

1. Аналитическая сторона вопроса: оценка размера группы и продолжительности проведения теста;

2. Выбор метрик исходя из бизнес-задачи:

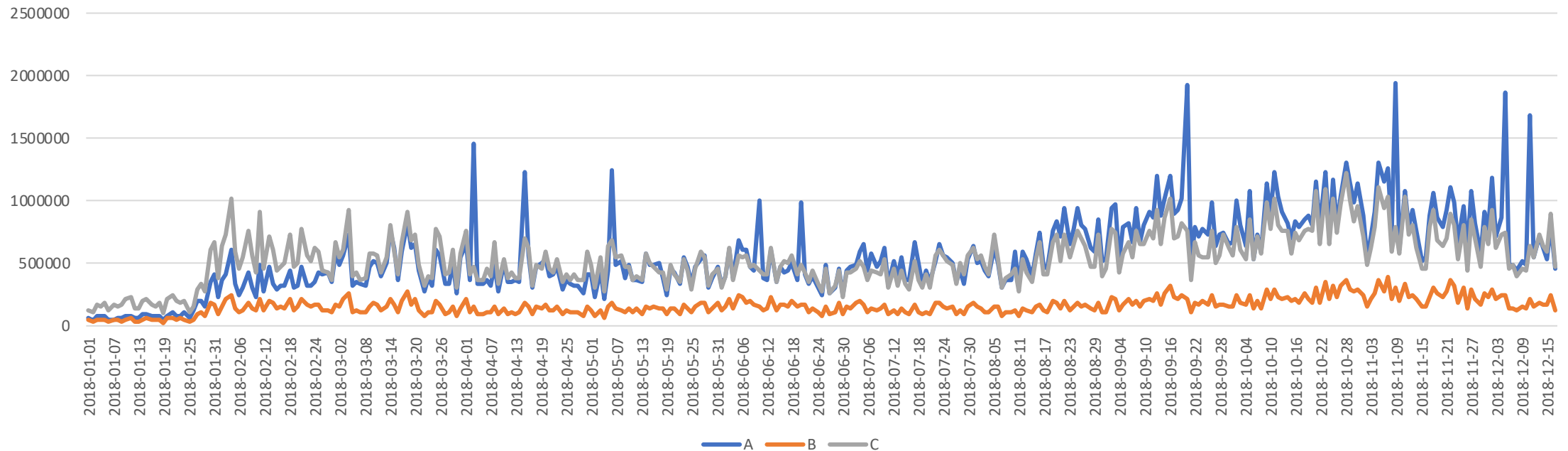
- churn-rate, retention;
- конверсионные метрики:
 - 1) из перехода на сайт – переход в корзину, в листинг, etc.;
 - 2) из показа карточки товара в listing'е в просмотр карточки;
 - 3) из просмотра карточки товара – в просмотр телефона магазина продавца или звонок, добавление в корзину, оплату товара из корзины, etc.
- относительные (на пользователя – поюзерные, на bucket – побакетные)
- абсолютные метрики (количество пользователей, перешедших на сайт, купивших услугу; количество купленных услуг, суммарная выручка)

3. Техническая сторона вопроса:

- формирование preliminary-TЗ (дизайнерам и разработчикам), принципы разбиения пользователей на bucket'ы, составление схемы логирования;

3. Запуск эксперимента

- Как проверяем, что наш эксперимент идёт в штатном режиме (основная идея - аналитик держит руку на пульсе и проверяет за группами тестирования и разработки);



4. Оценка полученных результатов

- Теоретический базис: основные понятия
- Виды выборок и А/В-экспериментов
- Основные критерии и границы их применения
- Оценка размера выборки
- Методы приведения выборок к нормальному виду

Виды выборок

- **Связанные** (зависимые, парные) – каждое наблюдение одной выборки неразрывно связано (находится в паре) с одним из наблюдений другой выборки;
- **Несвязанные** (независимые) – выборки, в которые объекты исследования набирались независимо друг от друга;

Типы проводимых экспериментов

- Повторные измерения;
- Множественные тесты;
- А/А-тесты

Основные понятия

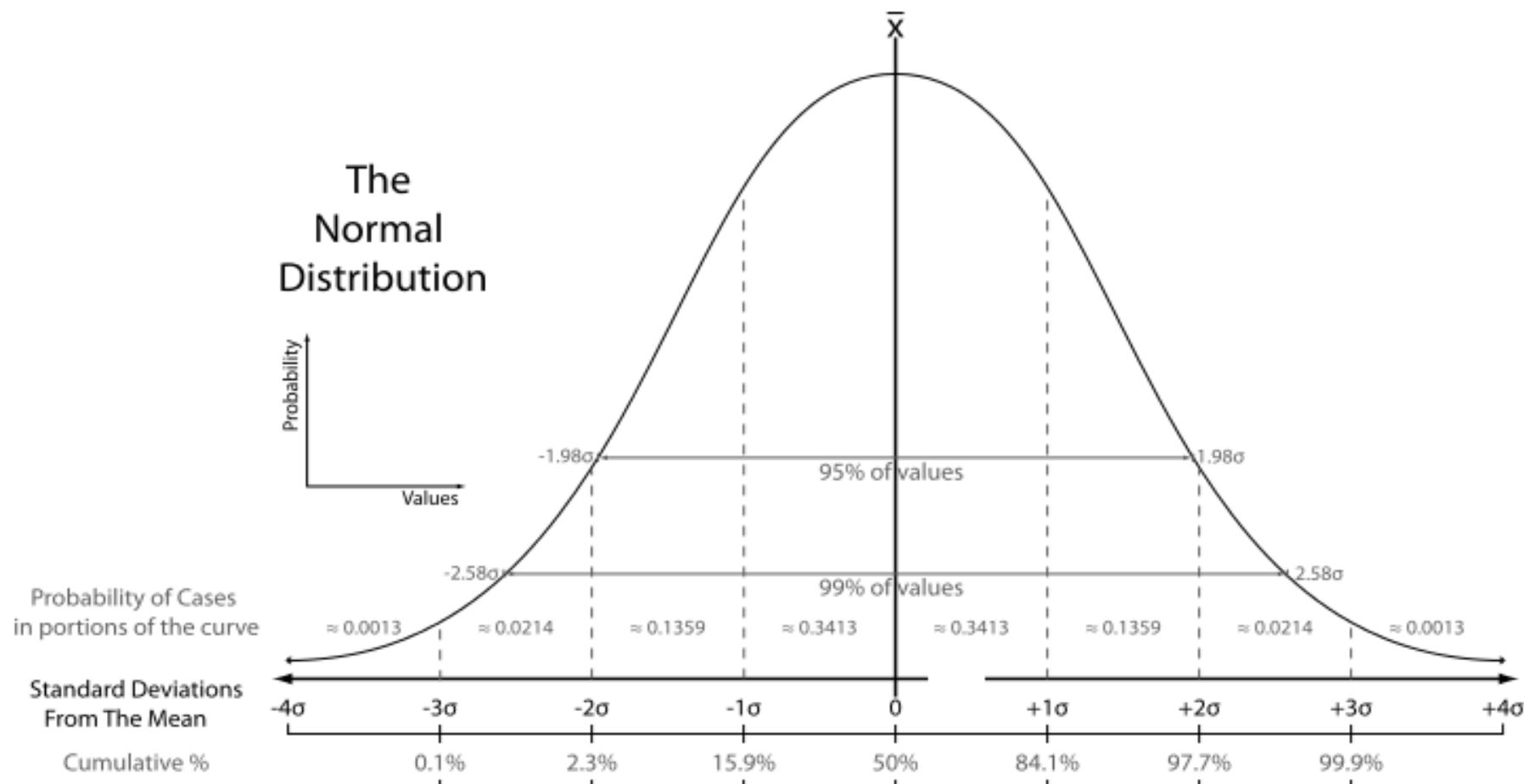


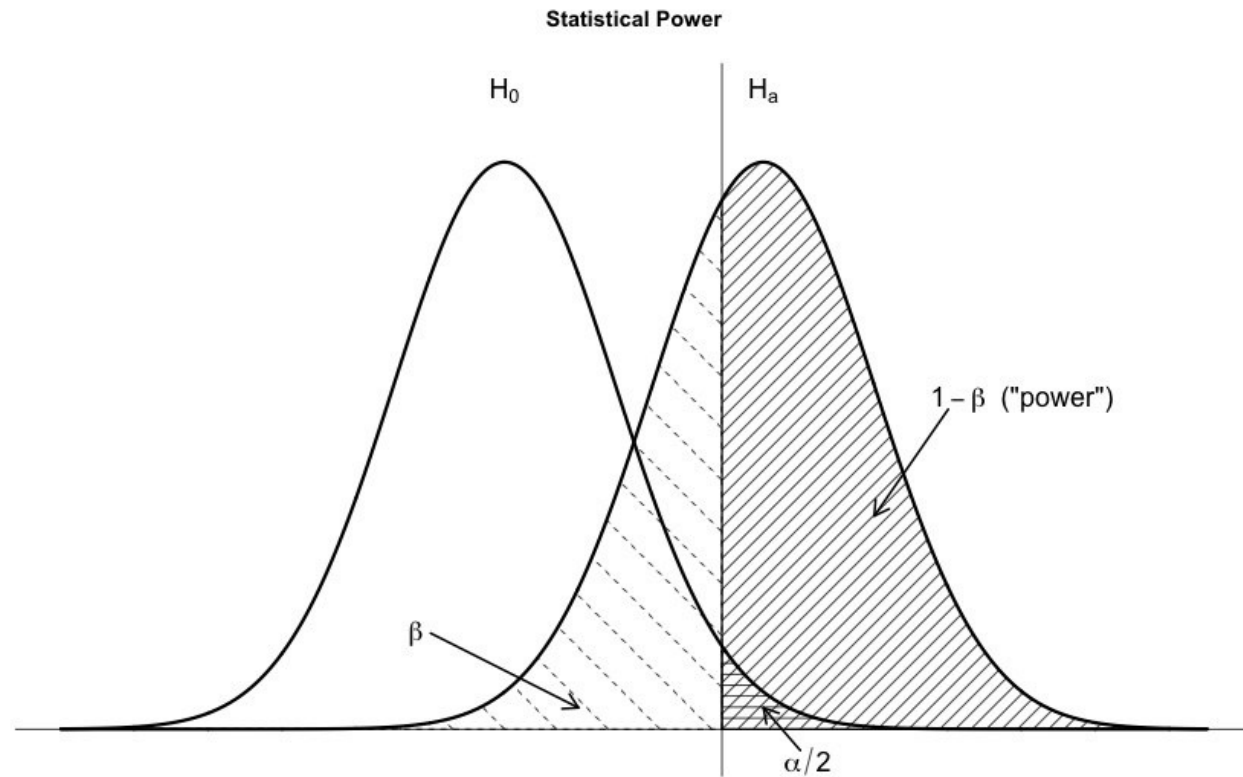
Рис. 1. Гистограмма нормального распределения

- **H₀** – основная гипотеза (о сходстве):
 $\mu(A) == \mu(B);$
- **H₁** – альтернативная гипотеза (о различии):
 $\mu(A) != \mu(B);$

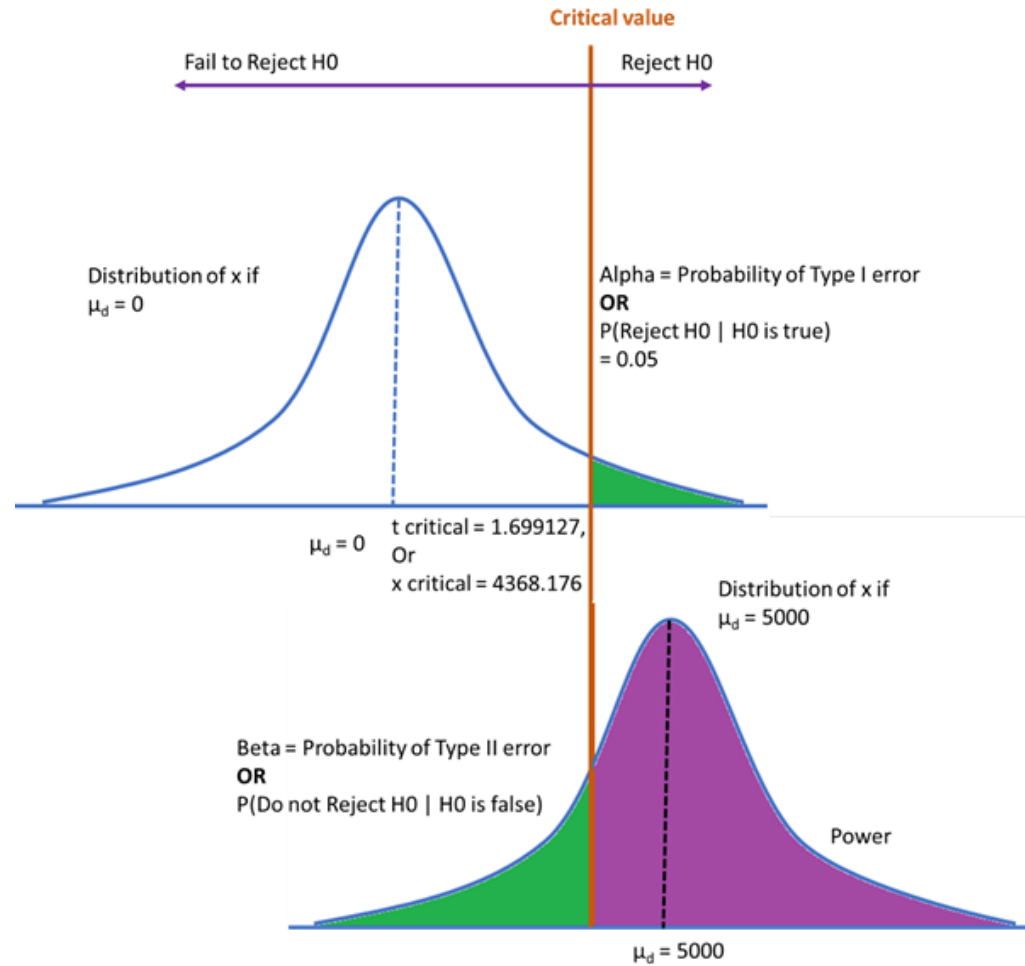
Виды гипотез

Результат применения критерия	Верная гипотеза		
		H0	H1
	H0	H0 верно принята (true positive)	H0 неверно принята (ошибка второго рода – «пропуск цели», false positive)
	H1	H0 неверно отвергнута (ошибка второго рода – «ложная тревога», false negative)	H0 верно отвергнута (true negative)

Уровень значимости и мощность критерия на гистограмме



Принцип принятия и отвержения гипотезы



Виды статистических критериев (I)

- **Параметрические** – основаны на конкретном типе распределения:

- 1) Т-Критерий Стьюдента;
- 2) Z-критерий Фишера;
- 3) F-критерий Фишера;
- 4) χ^2 -критерий Пирсона;

- **Непараметрические** - не базируется на предположении о типе распределения генеральной совокупности и не использует параметры этой совокупности:

- 1) Т-Критерий Уилкоксона;
- 2) U-Критерий Манна-Уитни;
- 3) Критерий Колмогорова;
- 4) Q-Критерий Розенбаума

Виды статистических критериев (II)

- Критерии согласия:

- 1) Критерий Колмогорова-Смирнова;
- 2) Критерий хи-квадрат (Пирсона);
- 3) Критерий Шапиро-Уилкса;

- Критерии сдвига (проверка равенства групп):

- 1) Критерий Стьюдента;
- 2) Критерий Уилкоксона;
- 3) Критерий Манна-Уитни;

- Критерии однородности (например, проверки на равенство дисперсий):

- 1) Критерий Барлетта;
- 2) Критерий Левенна;

Основные статистики

- Выборочное среднее:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + \dots + x_n)$$

- Смещённая оценка дисперсии:

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Среднеквадратическое отклонение
(несмещённая оценка дисперсии):

$$S_0 = \sqrt{\frac{n}{n-1} S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Математическое ожидание:

$$M[X] = \sum_{i=1}^{\infty} x_i p_i$$

Т-Критерий Стьюдента

Проверяется равенство средних в 2-х выборках;

- **Границы применимости:**

- 1) Равенство дисперсий;
- 2) Выборки имеют вид нормального распределения;
- 3) Неизвестна дисперсия генеральной совокупности;
- 4) Размер выборки < 30 (для малых выборок)

- **Статистика:**

- 1) одновыборочный критерий:

$$H_0 : E(X) = m$$

$$t = \frac{\bar{X} - m}{s_X / \sqrt{n}}$$

- 2) критерий для независимых выборок:

$$H_0 : M_1 = M_2$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Оценка размера группы для t-test'a

- Необходимая мощность исследования **(1 – β)** :

Мощность = $P(\text{отвергнуть } H_0 \mid H_1 \text{ is true})$

Мощность = $1 - P(\text{принять } H_0 \mid H_0 \text{ is false})$

Стандартная: 80%, **β = 0.8**, **$Z_{(1-\beta)} = 0.8416$**

- Необходимый уровень статистической значимости **α**:

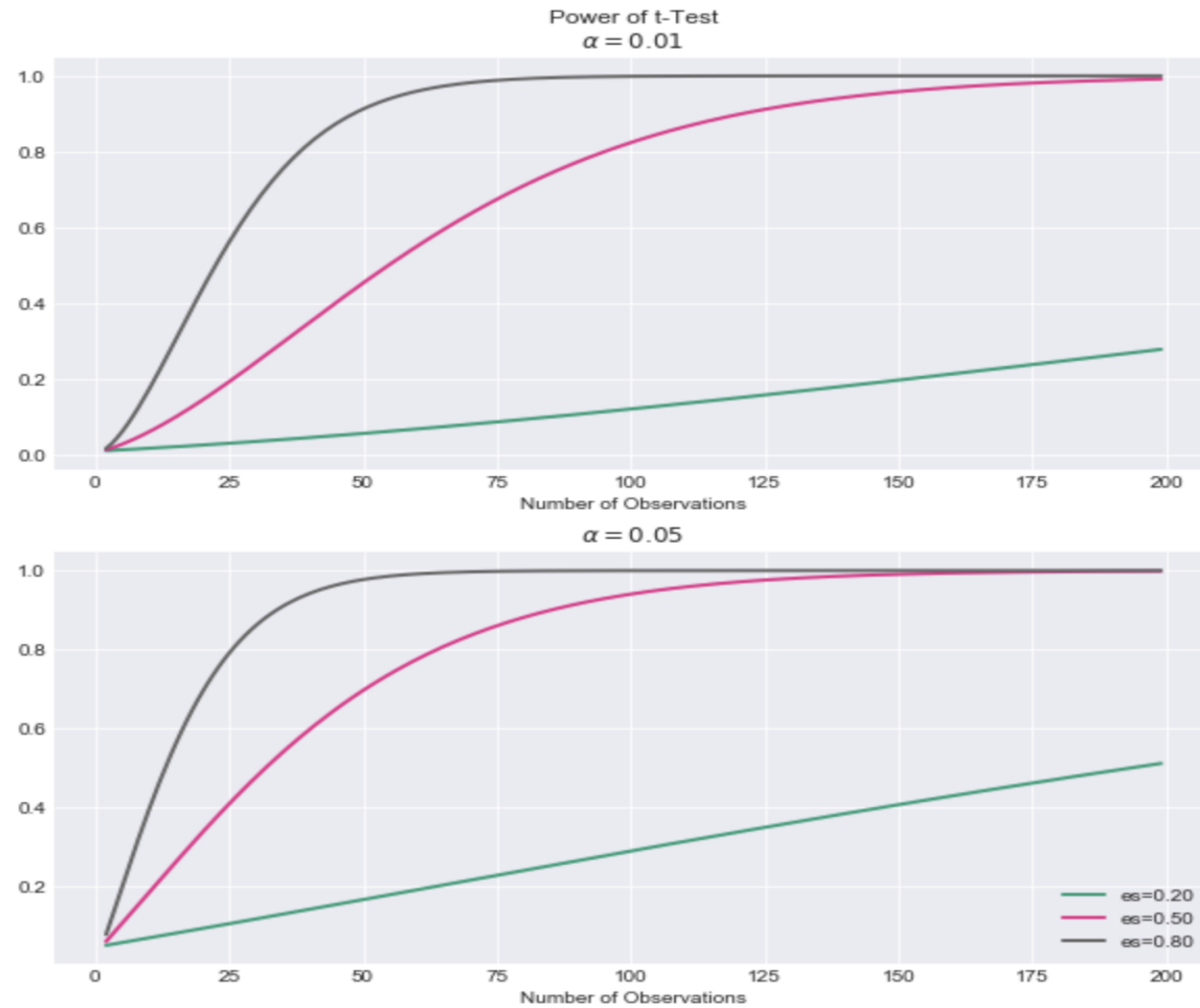
Стандартная: 5%, **α = 0.05**, **$Z_\alpha = 1.96$**

- **Размер эффекта Δ**: наша оценка – какую дельту между группами мы хотим увидеть?

- **Формула:**

$$n = \frac{2(Z_\alpha + Z_{1-\beta})^2 \sigma^2}{\Delta^2}$$

Оценка размера группы для t-test'a



Z-Критерий Фишера

Проверяется равенство средних в 2-х выборках;

- **Границы применимости:**

- 1) Равенство дисперсий;
- 2) Выборки имеют вид нормального распределения;
- 3) Известна дисперсия генеральной совокупности;
- 4) Размер выборки > 30 (для больших выборок)

- **Статистика:** $H_0 : M_x = m$
$$z_{\bar{X}} = \frac{\bar{X} - m_{H_0}}{\sigma / \sqrt{n}}$$

Методы приведения распределения к нормальному: ЦПТ и метод подгрупп

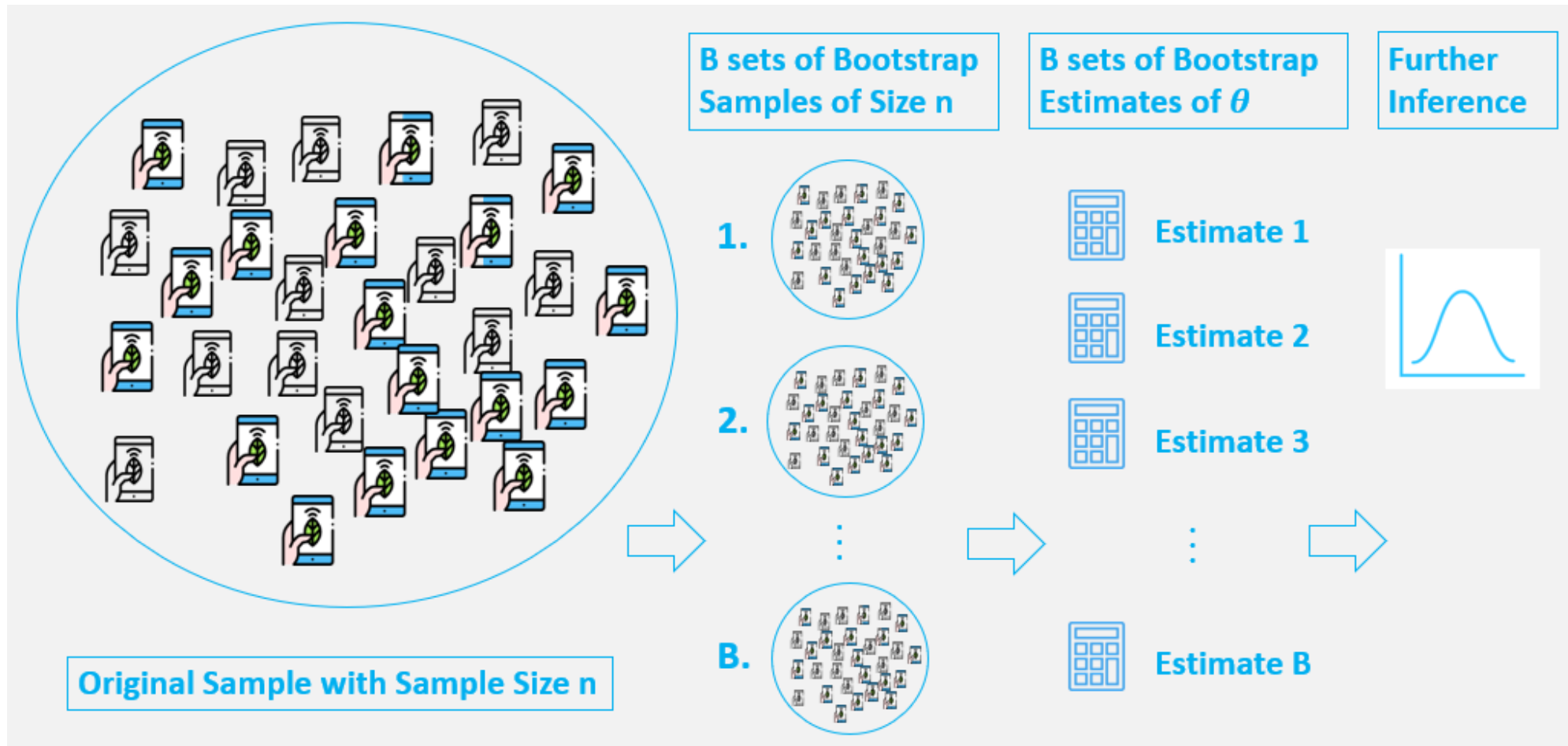
- **Центральная предельная теорема:**

”Сумма **большого** количества **независимых случайных** величин, имеющих примерно **одинаковые** масштабы (ни одно из слагаемых не вносит в сумму определяющего вклада), имеет распределение, близкое к **нормальному**.”

- **Метод sub-bucket’ов:**

Суть: разбить существующие **группы** (bucket’ы) на **подгруппы** (sub-bucket’ы) и, соответственно, взяв **средние значения** по подгруппам для каждого из bucket’ов, мы можем получить распределение, близкое к нормальному.

Методы приведения распределения к нормальному: Bootstrap



Методы приведения распределения к нормальному

- **Метод bootstrap:**

Суть: случайным образом, с повторениями, выбираются размещения из генеральной совокупности (исходной выборки) с возвращением. На выходе формируются средние (или медианы, суммы, etc.) от средних для каждой из сформированных подвыборок.

Метод реализован в библиотеке **Facebook Bootstrapped:**

<https://github.com/facebookincubator/bootstrapped>

- **Практическое задание:**

https://github.com/DilemmaLab/hse/blob/master/ab_bootstraping.py

Дополнительная информация:

- **Coursera:** курс “Построение выводов по данным”:
 - <https://ru.coursera.org/learn/stats-for-data-analysis>
- **Statistica.ru:** теория:
 - <http://statistica.ru/theory/proverka-gipotez/>
- **MachineLearning.ru:** теория:
 - http://www.machinelearning.ru/wiki/index.php?title=Проверка_статистических_гипотез
- **TowardsDataScience.com:** статьи с практическими примерами:
 - <https://towardsdatascience.com/how-to-use-python-to-figure-out-sample-sizes-for-your-study-871f0a76a19c>
 - <https://towardsdatascience.com/introduction-to-power-analysis-in-python-e7b748dfa26>