Яндекс

Skoltech
Skolkovo Institute of Science and Technology

Проверка гипотез. Часть 2

Лемма Неймана-Пирсона. Критерий отношения правдоподобия. Введение в A/B-тестирование. Критерий последовательного отношения правдоподобия. Непараметрические критерии.

ПМИ ФКН ВШЭ, 13 октября 2018 г.

Алексей Артемов[1,2]

[1] Сколтех [2] НИУ ВШЭ

Содержание лекции

- Лемма Неймана-Пирсона.
- Критерий отношения правдоподобия.
- Функции штрафа и основы теории принятия статистических решений.
- Введение в A/B-тестирование.
- Критерий последовательного отношения правдоподобия.
- Непараметрические критерии.

Лемма Неймана-Пирсона

Показательный пример

- Пусть $X_i \sim \mathcal{N}(\theta, \sigma^2), i = 1, \ldots, n$, а нулевая гипотеза заключается в том, что $\theta = \theta_0$.
- Назовите достаточную статистику в этой задаче.

Показательный пример

- Пусть $X_i \sim \mathcal{N}(\theta, \sigma^2), i = 1, \ldots, n$, а нулевая гипотеза заключается в том, что $\theta = \theta_0$.
- Назовите достаточную статистику в этой задаче.
- Достаточная статистика: $T(\mathbf{X}^\ell) = \overline{X}_n$.
- Какое распределение $T(\mathbf{X}^\ell)$, когда верна $\mathbb{H}_0$?

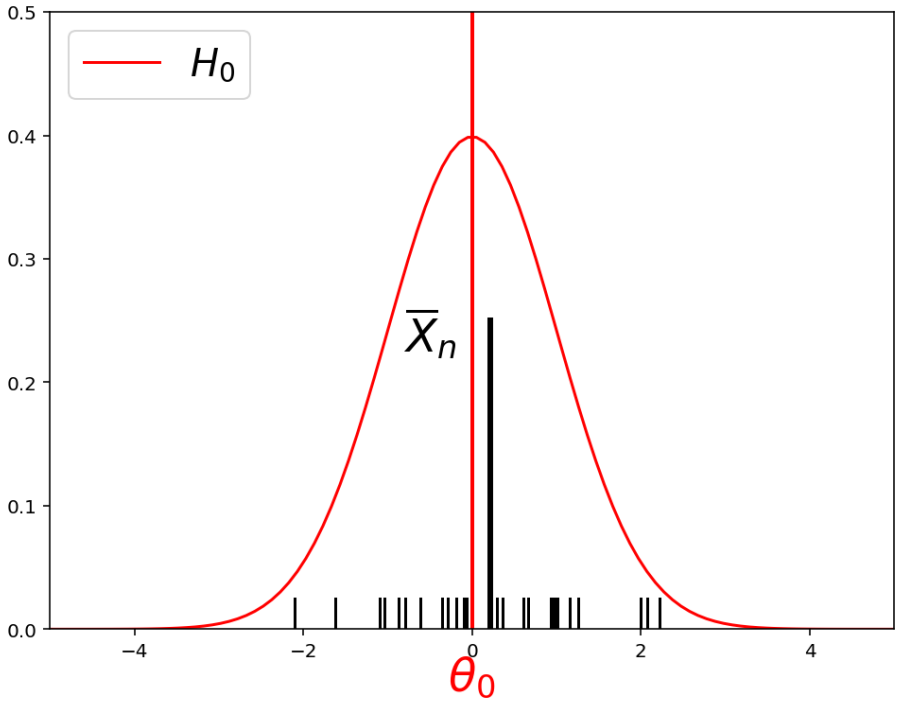
Показательный пример

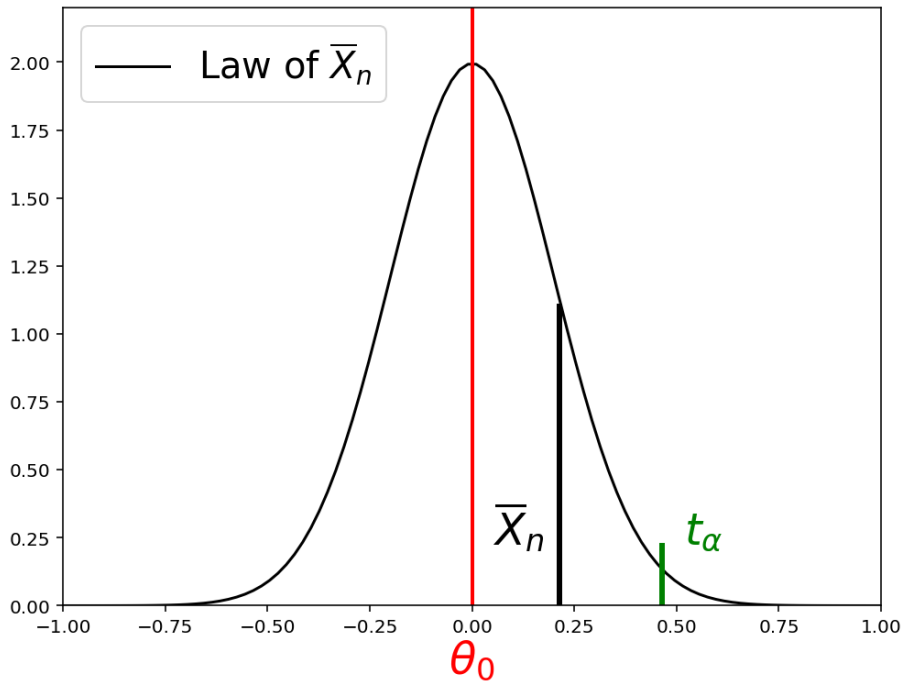
- Пусть $X_i \sim \mathcal{N}(\theta, \sigma^2), i = 1, \ldots, n$, а нулевая гипотеза заключается в том, что $\theta = \theta_0$.
- Назовите достаточную статистику в этой задаче.
- Достаточная статистика: $T(\mathbf{X}^\ell) = \overline{X}_n$.
- Какое распределение $T(\mathbf{X}^\ell)$, когда верна $\mathbb{H}_0$?
- $T(\mathbf{X}^\ell) \sim \mathcal{N}(\theta_0, \sigma^2/n)$
- Какова критическая область? (Когда отклоняем $\mathbb{H}_0$?)

Показательный пример

- Пусть $X_i \sim \mathcal{N}(\theta, \sigma^2), i = 1, \ldots, n$, а нулевая гипотеза заключается в том, что $\theta = \theta_0$.
- Назовите достаточную статистику в этой задаче.
- Достаточная статистика: $T(\mathbf{X}^\ell) = \overline{X}_n$.
- Какое распределение $T(\mathbf{X}^\ell)$, когда верна $\mathbb{H}_0$?
- $T(\mathbf{X}^\ell) \sim \mathcal{N}(\theta_0, \sigma^2/n)$
- Какова критическая область? (Когда отклоняем $\mathbb{H}_0$?)
- Критическая область $\mathcal{R}_\alpha = [t_\alpha, \infty)$, т.е. $T(\mathbf{X}^\ell) \geqslant t_\alpha$.

Показательный пример





Показательный пример

› Подсчитайте вероятность ложной тревоги в этой задаче.

Показательный пример

› Подсчитайте вероятность ложной тревоги в этой задаче.

$$\begin{aligned} \alpha &= \mathrm{P}_{\theta_0}\left(\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sigma} \geqslant \frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right) = \\ &= 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right). \end{aligned}$$

› Как выбрать t_α, чтобы $\alpha \leqslant \alpha_0$?

Показательный пример

› Подсчитайте вероятность ложной тревоги в этой задаче.

$$\begin{aligned}\alpha &= \mathrm{P}_{\theta_0}\left(\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sigma} \geqslant \frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right) = \\ &= 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right).\end{aligned}$$

› Как выбрать t_α, чтобы $\alpha \leqslant \alpha_0$?

$$t_{\alpha_0} = \theta_0 + \sigma x_{1-\alpha_0}/\sqrt{n}$$

Показательный пример

› Подсчитайте вероятность ложной тревоги в этой задаче.

$$\begin{aligned}\alpha &= \mathrm{P}_{\theta_0}\left(\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sigma} \geqslant \frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right) = \\ &= 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right).\end{aligned}$$

› Как выбрать t_α, чтобы $\alpha \leqslant \alpha_0$?

$$t_{\alpha_0} = \theta_0 + \sigma x_{1-\alpha_0}/\sqrt{n}$$

› Пусть на самом деле верна альтернатива $\mathbb{H}_1 : \theta = \theta_1$, причем $\theta_1 > \theta_0$. Какова вероятность ошибки 2-го рода?

Показательный пример

› Подсчитайте вероятность ложной тревоги в этой задаче.

$$\begin{aligned}\alpha &= \mathrm{P}_{\theta_0}\left(\frac{\sqrt{n}(\overline{X}_n - \theta_0)}{\sigma} \geqslant \frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right) = \\ &= 1 - \Phi\left(\frac{\sqrt{n}(t_\alpha - \theta_0)}{\sigma}\right).\end{aligned}$$

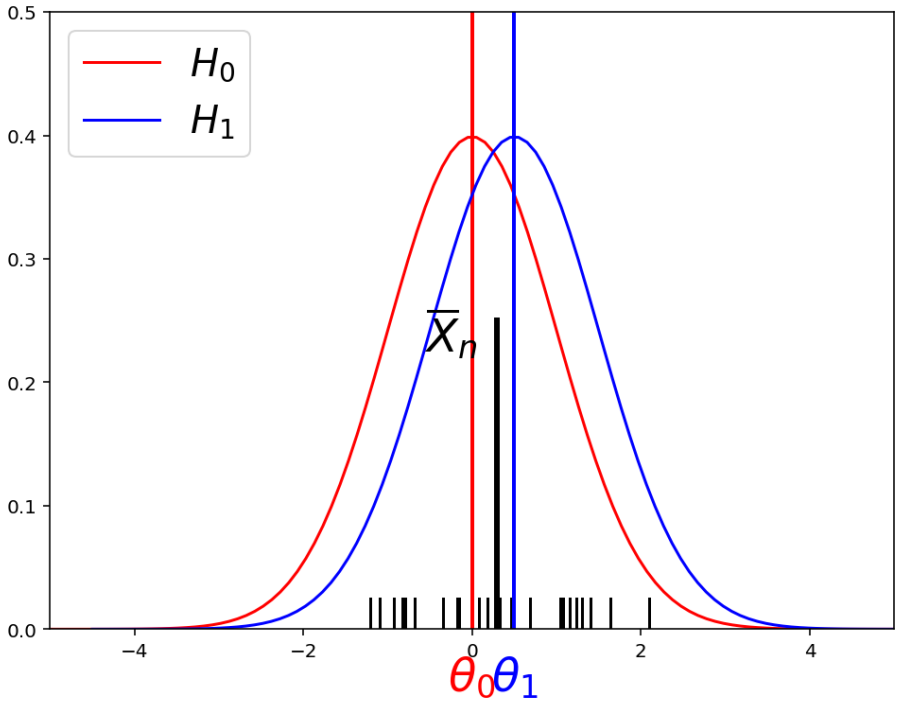
› Как выбрать t_α, чтобы $\alpha \leqslant \alpha_0$?

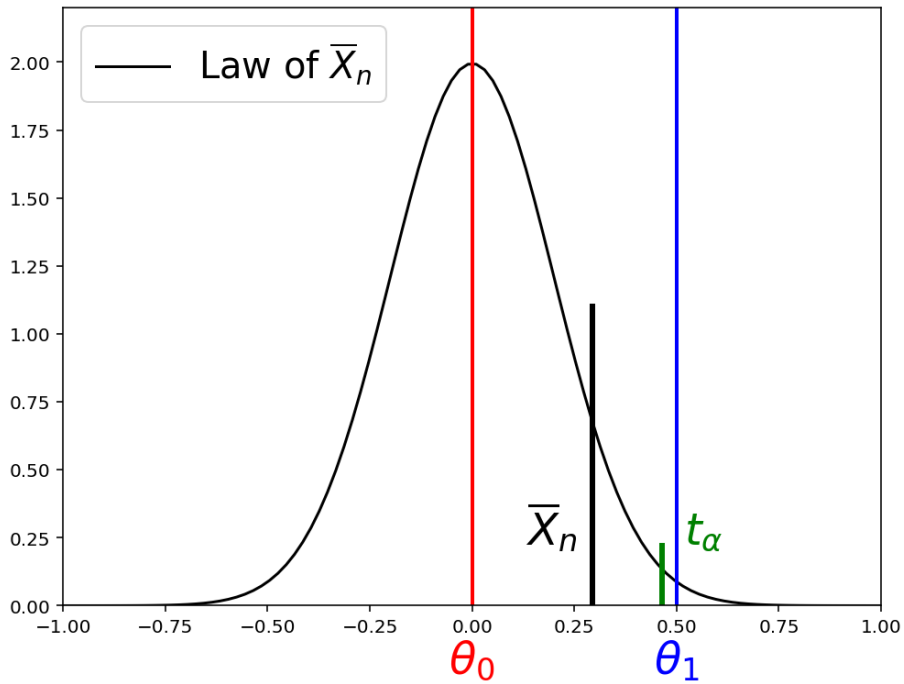
$$t_{\alpha_0} = \theta_0 + \sigma x_{1-\alpha_0}/\sqrt{n}$$

› Пусть на самом деле верна альтернатива $\mathbb{H}_1 : \theta = \theta_1$, причем $\theta_1 > \theta_0$. Какова вероятность ошибки 2-го рода?

$$\beta = \mathrm{P}_{\theta_1}\left(\overline{X}_n < t_{\alpha_0}\right) = \Phi\left(x_{1-\alpha_0} - \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sigma}\right).$$

Показательный пример



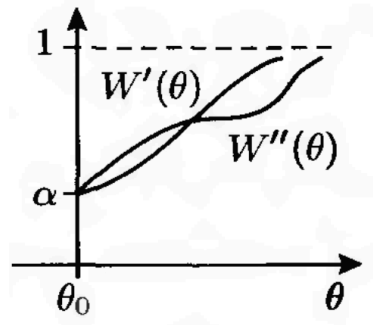


Сравнение двух критериев

- Критерии на самом деле задаются критическими множествами
- Пусть имеется два критерия, заданных множествами $\mathcal{R}'_\alpha$ и $\mathcal{R}''_\alpha$. **Какой выбрать?**

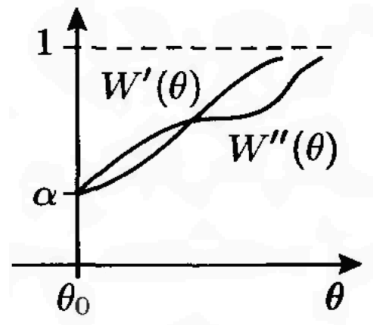
Сравнение двух критериев

- Критерии на самом деле задаются критическими множествами
- Пусть имеется два критерия, заданных множествами $\mathcal{R}'_\alpha$ и $\mathcal{R}''_\alpha$. **Какой выбрать?**
- **Сложная альтернатива:** необходимо сравнивать функции мощности $W'(\theta)$ и $W''(\theta)$



Сравнение двух критериев

- Критерии на самом деле задаются критическими множествами
- Пусть имеется два критерия, заданных множествами $\mathcal{R}'_\alpha$ и $\mathcal{R}''_\alpha$. **Какой выбрать?**
- **Сложная альтернатива:** необходимо сравнивать функции мощности $W'(\theta)$ и $W''(\theta)$



- **Простая альтернатива:** существует наиболее мощный критерий (Неймана-Пирсона).
- Идея: при заданной (достаточно малой) вероятности ошибки 1-го рода α постараться уменьшить вероятность ошибки 2-го рода β насколько возможно за счет подбора критического множества $\mathcal{R}_\alpha$.

Сравнение двух критериев

- Пусть задана выборка $\mathbf{X}^\ell$
- Гипотеза $\mathbb{H}_0$ и альтернатива $\mathbb{H}_1$ порождают в выборочном пространстве $\mathbb{R}^\ell$ меры P_0 и P_1
- Таким образом, необходимо найти множество G такое, что $\mathrm{P}_0(G) \leqslant \alpha$ и $\mathrm{P}_1(G) \to \sup\limits_{G:\mathrm{P}_0(G)\leqslant\alpha} \mathrm{P}_1(G)$
- Рассмотрим систему вложенных множеств $G_c = \{\boldsymbol{x} \in \mathbb{R}^\ell : \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \geqslant c\}$
- Пусть $\varphi(c) = \mathrm{P}_0(G_c)$, тогда $\varphi(c)$ убывает с ростом c

Сравнение двух критериев

- На самом деле, $\varphi(c)$ убывает быстрее, чем $1/c$:

$$1 \geqslant \mathrm{P}_1(G_c) = \int\limits_{G_c} p_1(\boldsymbol{x})d\boldsymbol{x} \geqslant c \int\limits_{G_c} p_0(\boldsymbol{x})d\boldsymbol{x} = c\,\mathrm{P}_0(G_c) = c\varphi(c).$$

- Далее еще потребуем, чтобы плотности $p_1(\boldsymbol{x})$ и $p_0(\boldsymbol{x})$ были всюду положительны
- Дополнительно потребуем, чтобы
  $\forall \alpha \in (0,1) \quad \exists c = c_\alpha : \quad \varphi(c_\alpha) = \alpha.$

Лемма Неймана-Пирсона

Лемма (Неймана-Пирсона)

Наиболее мощный критерий уровня α задается критическим множест

$$G^* = G_{c_\alpha} = \left\{ \boldsymbol{x} \in \mathbb{R}^\ell : \frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} \geqslant c_\alpha \right\}$$

- Пусть G — критическое множество уровня α.
- Тогда $\mathrm{P}_0(G_c) \geqslant \alpha = \mathrm{P}_0(G_{c_\alpha})$
- Пусть $I(\boldsymbol{x})$ — индикатор G_c, $I^*(\boldsymbol{x})$ — индикатор G_{c_α}
- Функция

$$f(\boldsymbol{x}) = (I^*(\boldsymbol{x}) - I(\boldsymbol{x}))(p_1(\boldsymbol{x}) - c_\alpha p_0(\boldsymbol{x}))$$

неотрицательна при всех $\boldsymbol{x} \in \mathbb{R}^\ell$

Лемма Неймана-Пирсона

› Функция

$$f(\boldsymbol{x}) = (I^*(\boldsymbol{x}) - I(\boldsymbol{x}))(p_1(\boldsymbol{x}) - c_\alpha p_0(\boldsymbol{x}))$$

неотрицательна при всех $\boldsymbol{x} \in \mathbb{R}^\ell$

› Поэтому

$$\begin{aligned} 0 \;\leqslant\; & \int\limits_{\boldsymbol{x}\in\mathbb{R}^\ell} f(\boldsymbol{x})d\boldsymbol{x} = \int\limits_{\boldsymbol{x}\in\mathbb{R}^\ell} I^*(\boldsymbol{x})p_1(\boldsymbol{x})d\boldsymbol{x} - \int\limits_{\boldsymbol{x}\in\mathbb{R}^\ell} I(\boldsymbol{x})p_1(\boldsymbol{x})d\boldsymbol{x} - \\ - \;\; & c_\alpha\Big[\int\limits_{\boldsymbol{x}\in\mathbb{R}^\ell} I^*(\boldsymbol{x})p_0(\boldsymbol{x})d\boldsymbol{x} - \int\limits_{\boldsymbol{x}\in\mathbb{R}^\ell} I(\boldsymbol{x})p_0(\boldsymbol{x})d\boldsymbol{x}\Big] = \\ = \;\; & \mathrm{P}_1(G^*) - \mathrm{P}_1(G) - c_\alpha \underbrace{[\mathrm{P}_0(G^*) - \mathrm{P}_0(G)]}_{\geqslant 0}. \end{aligned}$$

Критерий Неймана-Пирсона: 2 простые гипотезы

Критерий Неймана-Пирсона

- $\mathbb{H}_0 : \theta = \theta_0$ vs. $\mathbb{H}_1 : \theta = \theta_1$
- Статистика Неймана-Пирсона:

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^{n} f(X_i; \theta_1)}{\prod_{i=1}^{n} f(X_i; \theta_0)}. \tag{1}$$

- Допустим, что $\mathbb{H}_0$ отвергается при $T > k$. Выберем k так, что $\mathrm{P}_{\theta_0}(T > k) = \alpha$.
- Тогда, критерий Неймана-Пирсона (на основе статистики (2)) будет иметь наибольшую мощность $W(\theta_1)$ среди всех критериев размера α.

Пример

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$, причем дисперсия σ^2 известна
- $\mathbb{H}_0 : \mu = \mu_0$ vs. $\mathbb{H}_1 : \mu = \mu_1$
- Статистика Неймана-Пирсона:

$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^{n} \mathcal{N}(X_i; \mu_1, \sigma^2)}{\prod_{i=1}^{n} \mathcal{N}(X_i; \mu_0, \sigma^2)}. \tag{2}$$

- Подсчитайте статистику критерия (упростите)

Пример

- $X_i \sim \mathcal{N}(\mu, \sigma^2)$, причем дисперсия σ^2 известна
- $\mathbb{H}_0 : \mu = \mu_0$ vs. $\mathbb{H}_1 : \mu = \mu_1$
- Статистика Неймана-Пирсона:

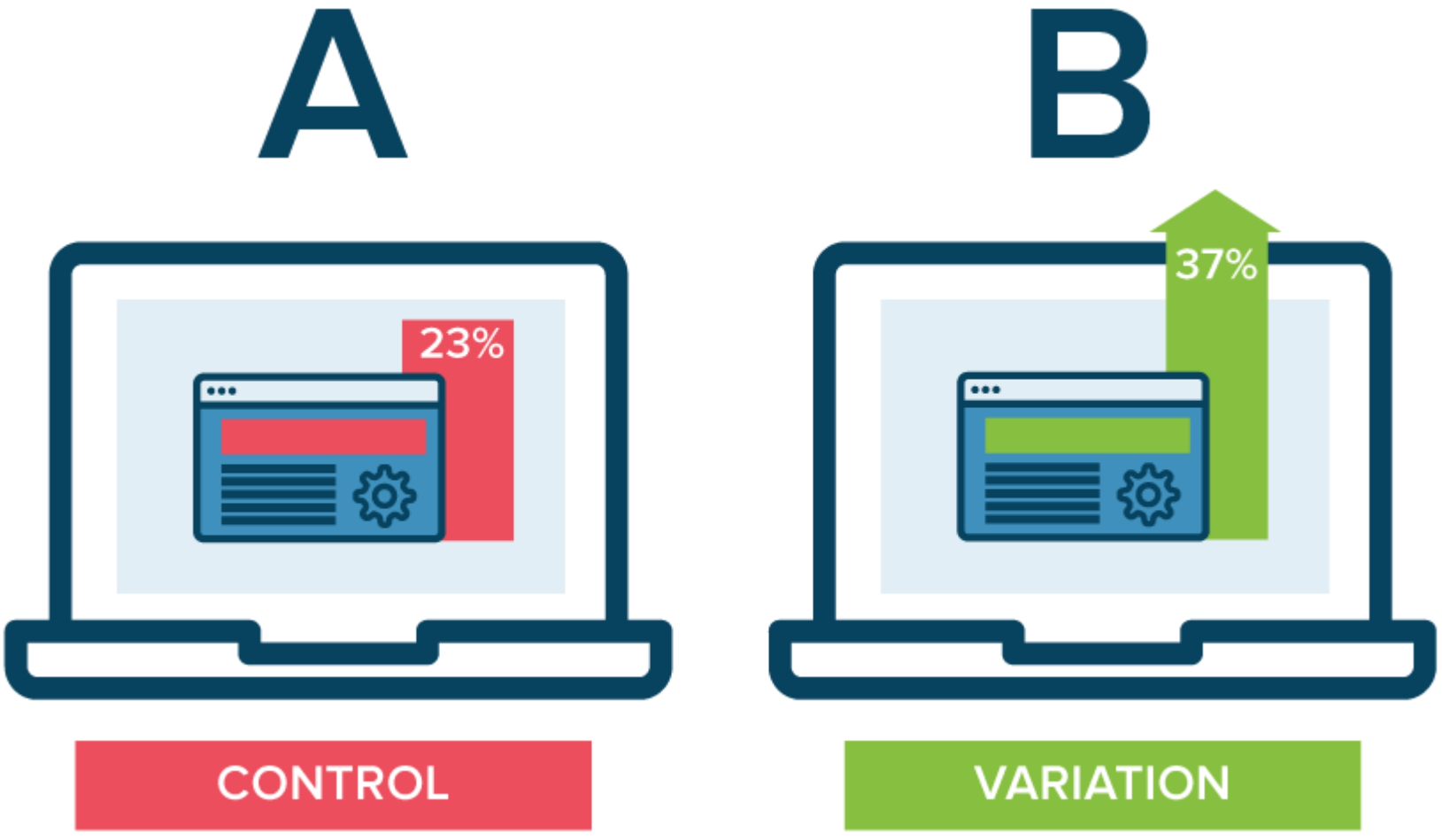
$$T = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} = \frac{\prod_{i=1}^{n} \mathcal{N}(X_i; \mu_1, \sigma^2)}{\prod_{i=1}^{n} \mathcal{N}(X_i; \mu_0, \sigma^2)}. \tag{2}$$

- Подсчитайте статистику критерия (упростите)
- Получается

$$\begin{aligned} T &= \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{\ell} \big[(X_i - \mu_1)^2 - (X_i - \mu_0)^2 \big] \Big] = \\ &= \exp\Big[\frac{n}{\sigma^2} (\mu_1 - \mu_0) \underbrace{\Big[\overline{X}_\ell - \frac{\mu_1 + \mu_0}{2} \Big]}_{\text{важен знак!}} \Big] \end{aligned}$$

Краткий экскурс в A/B тестирование

Идея A/B тестирования



Техника A/B тестирования

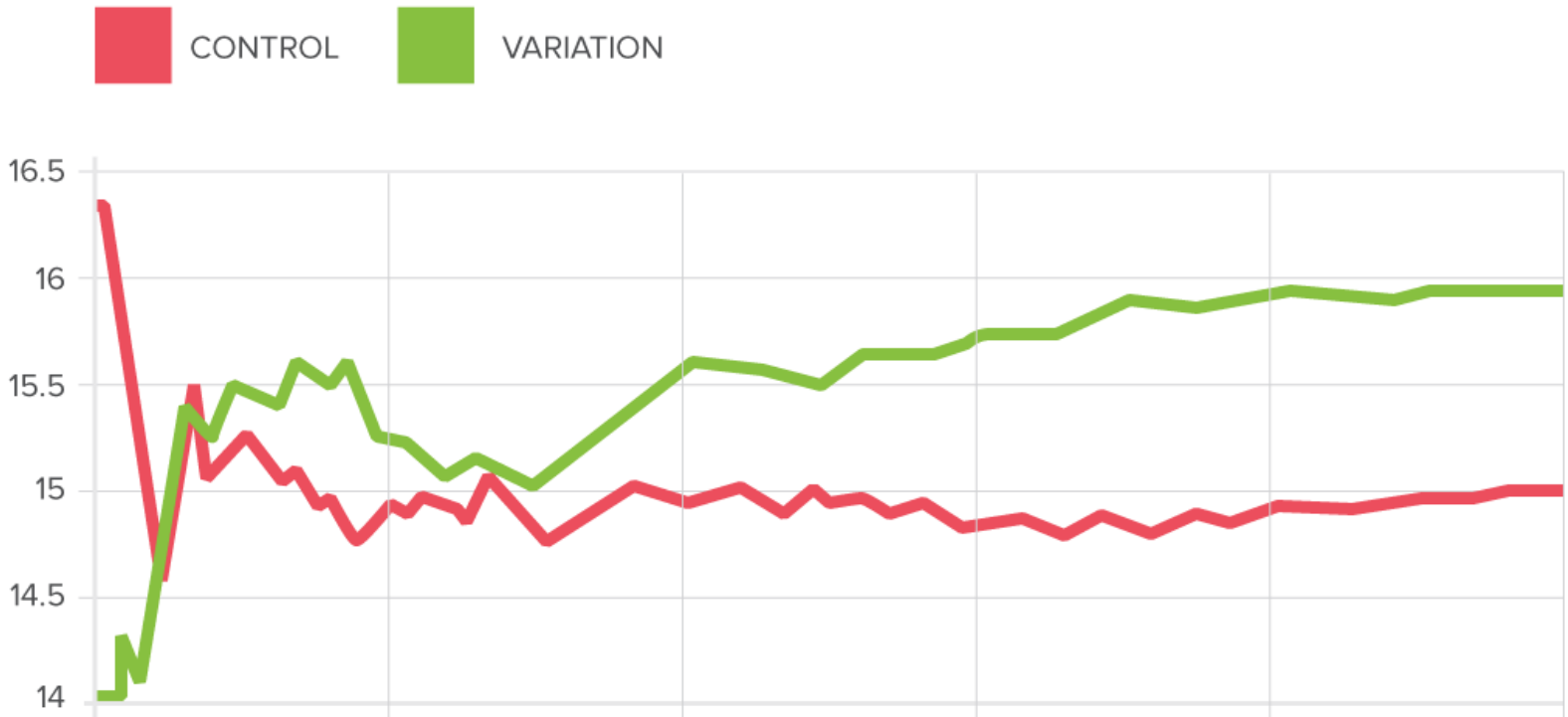
- Варианты разбиения
  - По пользователям (куки)
  - По визитам (сессии)
  - По действиям (запросы)
  - ...
- Типы экспериментов
  - явные (интерфейсы, функциональность)
  - неявные (ранжирование, персонализация)
  - сервисы / части сервисов / кросс- сервисные
  - улучшения / ухудшения / AA-тесты

Внедрения A/B тестирования

- Масштабы экспериментов:
  Яндекс @ 2017:
  - 4778 экспериментов,
  - $\sim$400 экспериментов одновременно
- Результаты:
  - Яндекс, 2014: 21% принимаются, 2017: 28% принимаются
  - Бинг, 201?: 30% принимаются
  - Гугл, 201?: 10% принимаются
  - В среднем: 20% принимаются, 50% непонятных

Техника A/B тестирования

- Метрики
  - абсолютные (клики, временные)
  - относительные (ctr, доля некликнутых)



Критерий Стьюдента (t-test)

Определение

Случайная величина имеет распределение Стьюдента (t-распределение) с k степенями свободы, если:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1+\frac{t^2}{k})^{\frac{k+1}{2}}}$$

При $k \to \infty$ t-распределение стремится к стандартному нормальному распределению. При $k = 1$ t-распределение совпадает с распределением Коши.

- t-критерий используют, когда распределение данных близко к нормальному, а размер выборки невелик.

Критерий Стьюдента (t-test)

Теорема (Критерий Стьюдента (t-test))

Пусть $X_1, ..., X_n \sim N(\mu, \sigma^2)$, где параметры (μ, σ^2) неизвестны.

$$\mathbb{H}_0 : \mu = \mu_0 \quad vs. \quad \mathbb{H}_1 : \mu \neq \mu_0$$

Обозначим через S_n^2 выборочную дисперсию. Тогда статистика критерия:

$$T = \frac{\sqrt{n}(\overline{X_n} - \mu_0)}{S_n}$$

Основная гипотеза отвергается, если $|T| > t_{n-1,\alpha/2}$, где $t_{n-1,\alpha/2}$ — квантиль распределения Стьюдента с $\mathbf{n-1}$ степенями свободы.

При больших n выполняется $T \sim \mathcal{N}(0, 1)$, то есть при больших n t-критерий эквивалентен критерию Вальда.

Критерий Стьюдента (t-test)

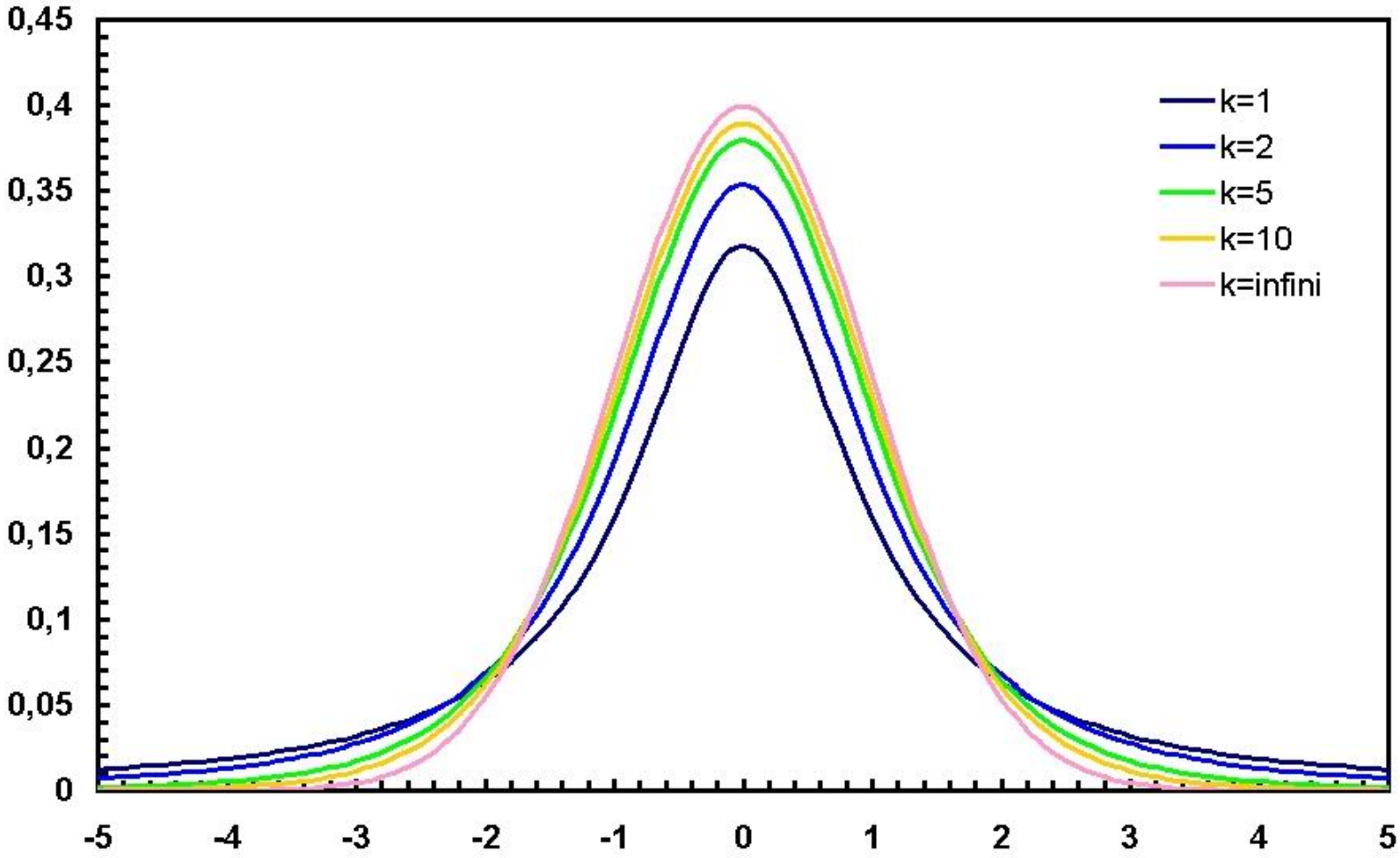


Рис.: https://ru.wikipedia.org/wiki/Распределение_Стьюдента

Критерий Стьюдента (t-test)

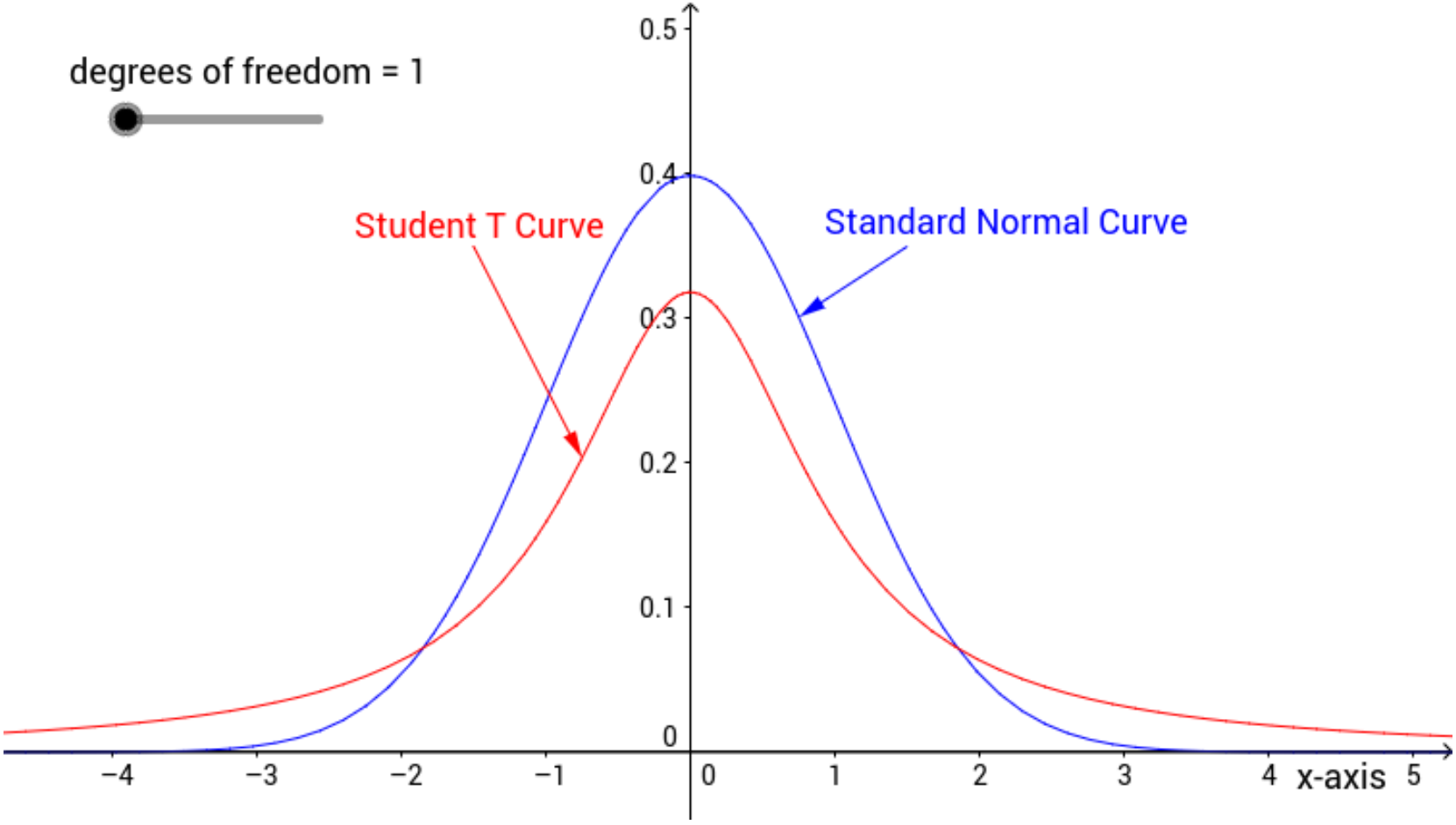


Рис.: http://tananyag.geomatech.hu/m/53882

Последовательный анализ

Мотивация

- В классической теории математической статистики предполагается, что наблюдения (данные) заранее известны
- Однако нужно ли заранее фиксировать размер выборки?
- В действительности, размер выборки можно определять в зависимости от уже поступивших наблюдений!
- Это приводит к последовательному тесту Вальда (последовательному критерию отношения правдоподобия, sequential probability ratio test, SPRT)
- Основная мотивация: **экономить данные** (до 50% экономии — доказательство на семинаре!)

Последовательный критерий

- Гипотеза $\mathbb{H}_0$: наблюдения X_i имеют плотность $p_0(x)$, $\mathbb{H}_1$: наблюдения X_i имеют плотность $p_1(x)$
- Рассмотрим $Z_i = \log \frac{p_0(X_i)}{p_1(X_i)}$ и блуждание $S_k = Z_1 + \ldots + Z_k$
- Например, если $p_0(x) = \mathcal{N}(\mu_0, \sigma^2)$, $p_1(x) = \mathcal{N}(\mu_1, \sigma^2)$, то

$$Z_i = \frac{\mu_1 - \mu_0}{\sigma^2}\left[\overline{X}_i - \frac{\mu_1 + \mu_0}{2}\right]$$

- Интуиция:

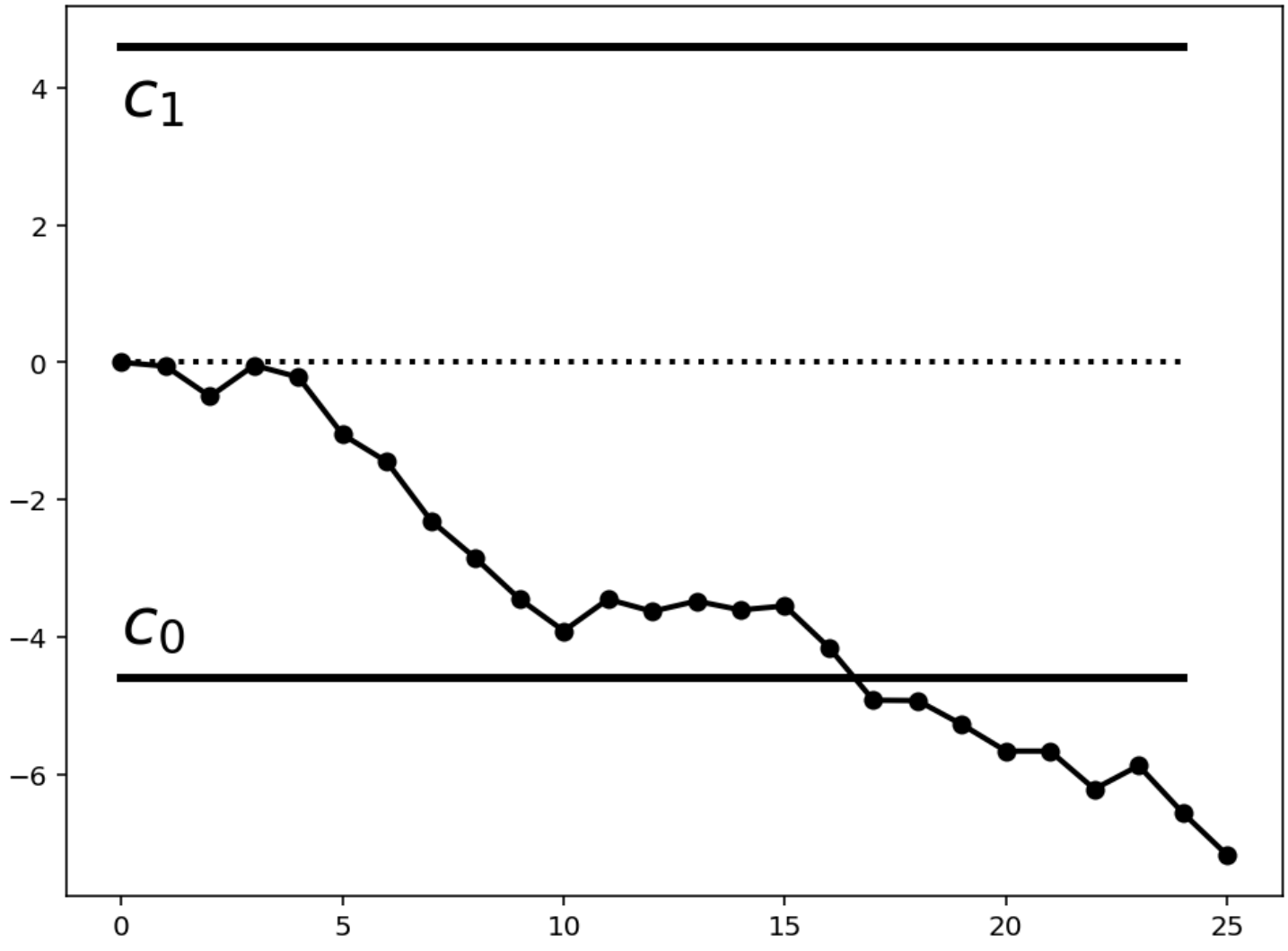
Последовательный критерий

- Гипотеза $\mathbb{H}_0$: наблюдения X_i имеют плотность $p_0(x)$, $\mathbb{H}_1$: наблюдения X_i имеют плотность $p_1(x)$
- Рассмотрим $Z_i = \log \frac{p_0(X_i)}{p_1(X_i)}$ и блуждание $S_k = Z_1 + \ldots + Z_k$
- Например, если $p_0(x) = \mathcal{N}(\mu_0, \sigma^2)$, $p_1(x) = \mathcal{N}(\mu_1, \sigma^2)$, то

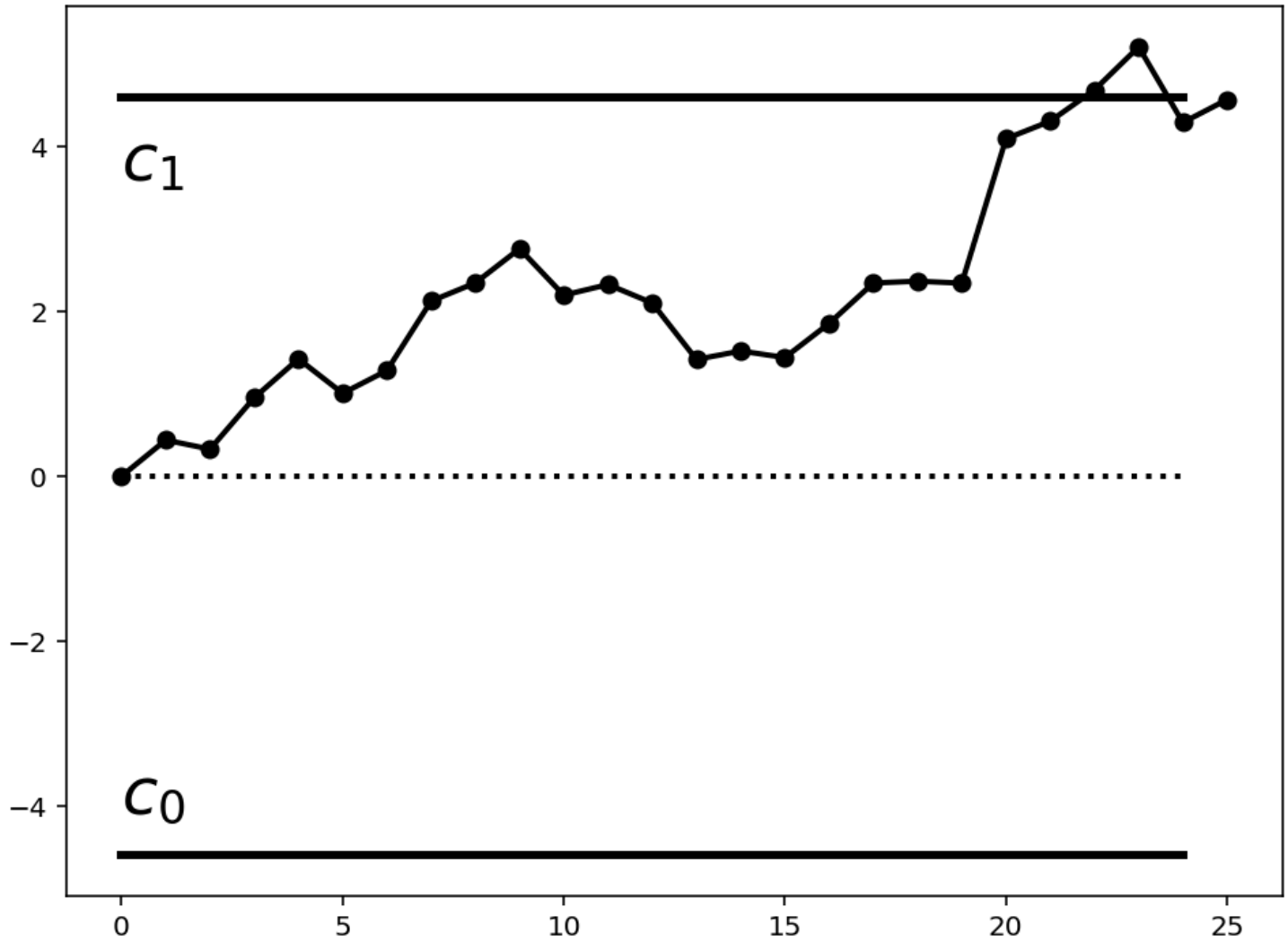
$$Z_i = \frac{\mu_1 - \mu_0}{\sigma^2}\left[\overline{X}_i - \frac{\mu_1 + \mu_0}{2}\right]$$

- Интуиция:
  - если верна $\mathbb{H}_0$ (сигнала нет), то в среднем $p_0(X_i) \geqslant p_1(X_i)$ (и тогда $Z_i \leqslant 0$) $\implies$ S_k убывает
  - если верна $\mathbb{H}_1$ (сигнал!), то в среднем $p_0(X_i) \leqslant p_1(X_i)$ (и тогда $Z_i \geqslant 0$) $\implies$ S_k растет

Пример: верна гипотеза



Пример: верна альтернатива



Последовательный критерий

- Последовательный тест:
  1. Наблюдаем $X_1, X_2, \ldots$ последовательно
  2. Вычисляем значения $S_k, k = 1, 2, \ldots$
  3. $S_k \geqslant c_1$: остановимся и отклоним $\mathbb{H}_0$
  4. $S_k \geqslant c_1$: остановимся и не отклоним $\mathbb{H}_0$
  5. $c_0 < S_k < c_1$: продолжим наблюдения
- Остается задать пределы роста S_k:

$$c_0 = \ln \frac{\beta'}{1-\alpha'}, \qquad c_0 = \ln \frac{1-\beta'}{\alpha'}, \qquad \alpha' + \beta' < 1$$

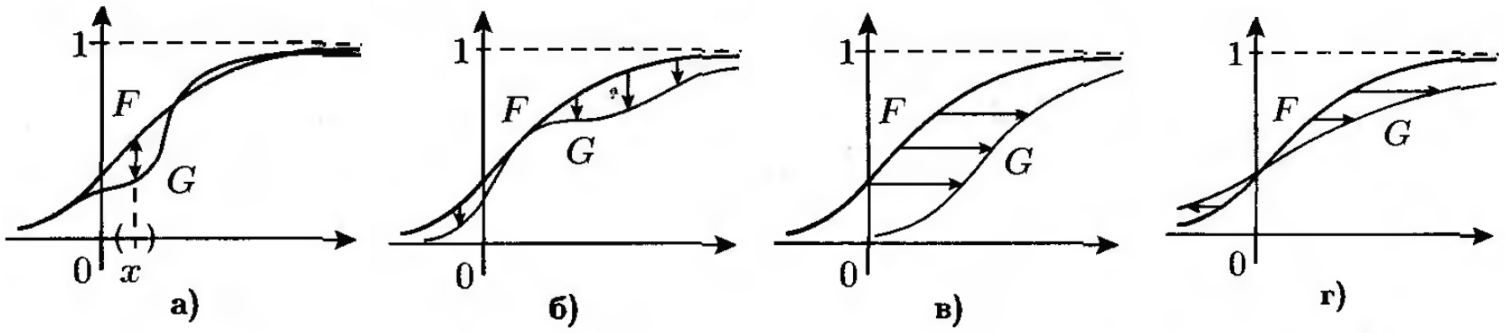
При этом для вероятностей α и β:

$$\alpha \leqslant \frac{\alpha'}{1-\beta'}, \qquad \beta \leqslant \frac{\beta'}{1-\alpha'}, \qquad \alpha + \beta < \alpha' + \beta'.$$

Непараметрические критерии

Альтернативы однородности

- Имеем две выборки $\mathbf{X}^n \sim F(x)$ и $\mathbf{Y}^m \sim G(x)$
- Гипотеза однородности $\mathbb{H}_0 : F(x) = G(x), x \in \mathbb{R}$



- Бывает важно уловить отклонения от $\mathbb{H}_0$ только определенного типа (наличие прироста Y_j по сравнению с X_i)
- Сужение типов альтернатив $\Longrightarrow$ более эффективные критерии
- Проверяем гипотезу однородности против альтернативы доминирования $\mathbb{H}_1$ (варианты б) и в))

Критерий ранговых сумм MWW

- Построим вариационный ряд из объединенной выборки $(X_1, \ldots, X_n, Y_1, \ldots, X_m)$
  - Верна $\mathbb{H}_0 \implies$ значения Y_j рассеяны по всему ряду
  - Иначе средний ранг значений Y_j относительно большой
- Обозначим S_j ранг порядковой статистики $Y_{(j)}$ в этом ряду
- Положим $V = S_1 + \cdots + S_m$
- Критическая область: $V \geqslant c$, где $c = \text{const}$
- **Большие выборки:** (Mann-Whitney-Wilcoxon, MWW)

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{X_i < Y_j} \to \mathcal{N}\left(\frac{nm}{2}, \frac{nm(n+m+1)}{12}\right)$$

Резюме лекции

- Критерий отношения правдоподобия: оптимален, если у вас простые гипотезы.
- A/B-тестирование — проверка гипотез в продакшене.
- Критерий последовательного отношения правдоподобия: снижение затрат на тестирование.
- Непараметрический ранговый критерий: рабочая лошадка АВТ (наряду с t-тестом!).