



> Конспект > 1 урок > Основы статистики и статистические критерии

> Оглавление

> Оглавление

> Методы принятия решения

> Тестирование гипотез

> Критическая область

> Этапы проверки статистических гипотез

> Важные статистические тесты

> Тест Стьюдента

> Тест Манна-Уитни

> Критерий Колмогорова

> Критерий отношения правдоподобия

> Бутстреп

> Доверительный интервал

> Бутстреп

> Проверка гипотезы о равенстве средних

> Нормальный интервал

> Интервал на основе процентилей

> Центральный интервал

> Как выбрать критерий

> Ошибки I и II рода

> Оценка ошибки I рода

> Оценка ошибки II рода

> p-value

> Распределение p-value

КУПЛЕНО НА
SKLADCHIK.COM

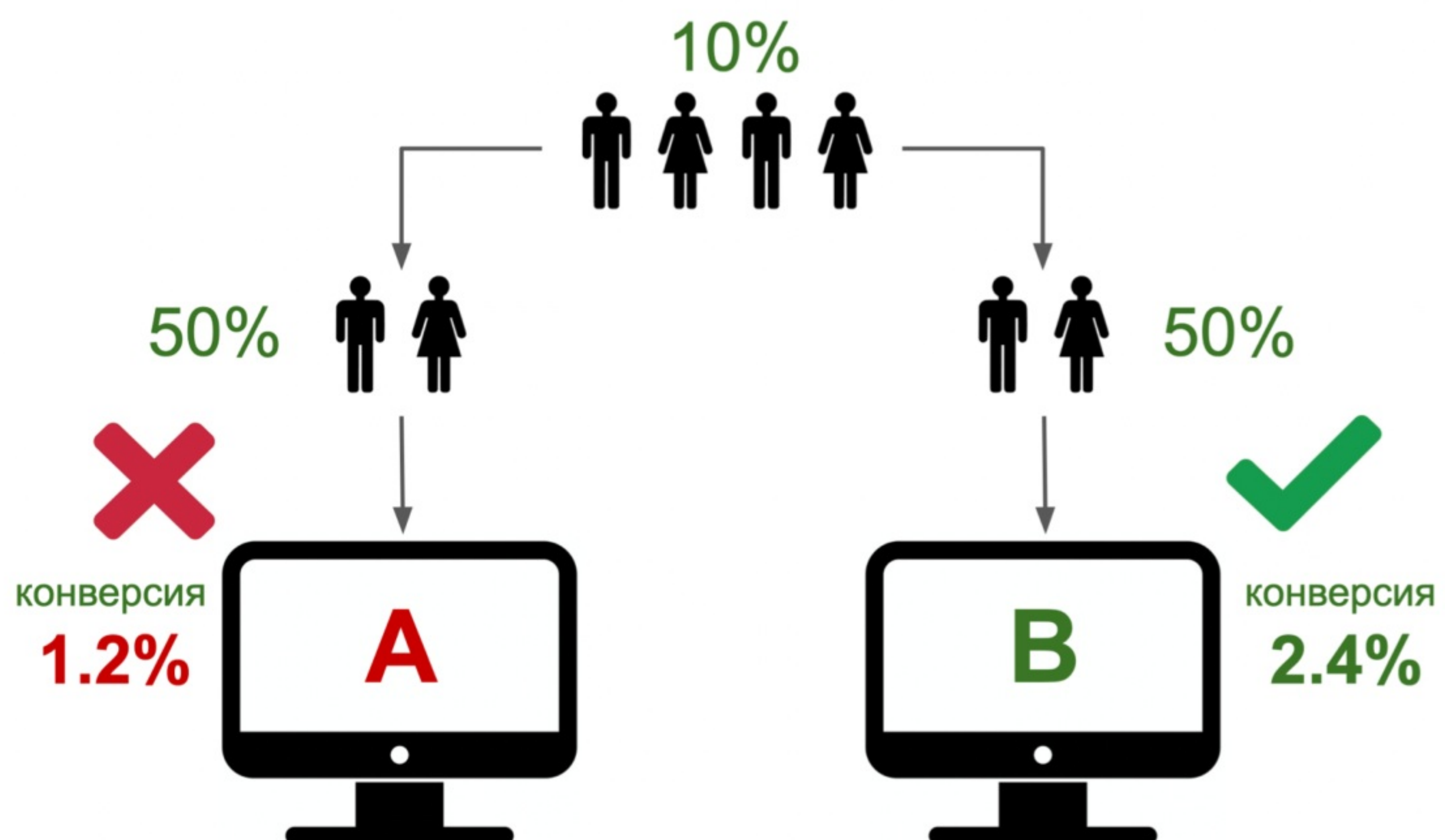
> Методы принятия решения

Как принимали решения раньше:

- Опрос пользователей;
- Сравнение метрики до и во время эксперимента;
- Мнение эксперта.

Определение

A/B-тестирование — метод, который позволяет на основе сравнения пилотной и контрольной групп оценивать изолированный эффект внедряемых изменений.



> Тестирование гипотез

Определение

Статистическая гипотеза — любое предположение о распределении и свойствах случайной величины.

Примеры:

- Несколько простых гипотез: $H_0 = \{F = F_0\}, H_1 = \{F = F_1\}$

- Простая основная и сложная альтернативная гипотезы: $H_0 = \{EX = EY\}$, $H_1 = \{EX \neq EY\}$
- Гипотеза независимости: $H_0 = \{P(X) = P(X|Y)\}$, $H_1 = \{H_0 \text{ неверна}\}$

Определение

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

Дана выборка $X_1, \dots, X_n \sim F$.

Хотим проверить простую гипотезу H_0 против сложной альтернативы H_1 .

Пусть можно задать функцию $t(X_n)$, обладающую свойствами:

1. Если H_0 верна, то $t(X_n) \Rightarrow G$, где G — непрерывное распределение.
2. Если H_0 неверна, то $|t(X_n)| \xrightarrow{P} \infty$ при $n \rightarrow \infty$.

Для СВ $Y \sim G$ определим постоянную C из равенства $\alpha = P(|Y| \geq C)$.

Тогда критерий:

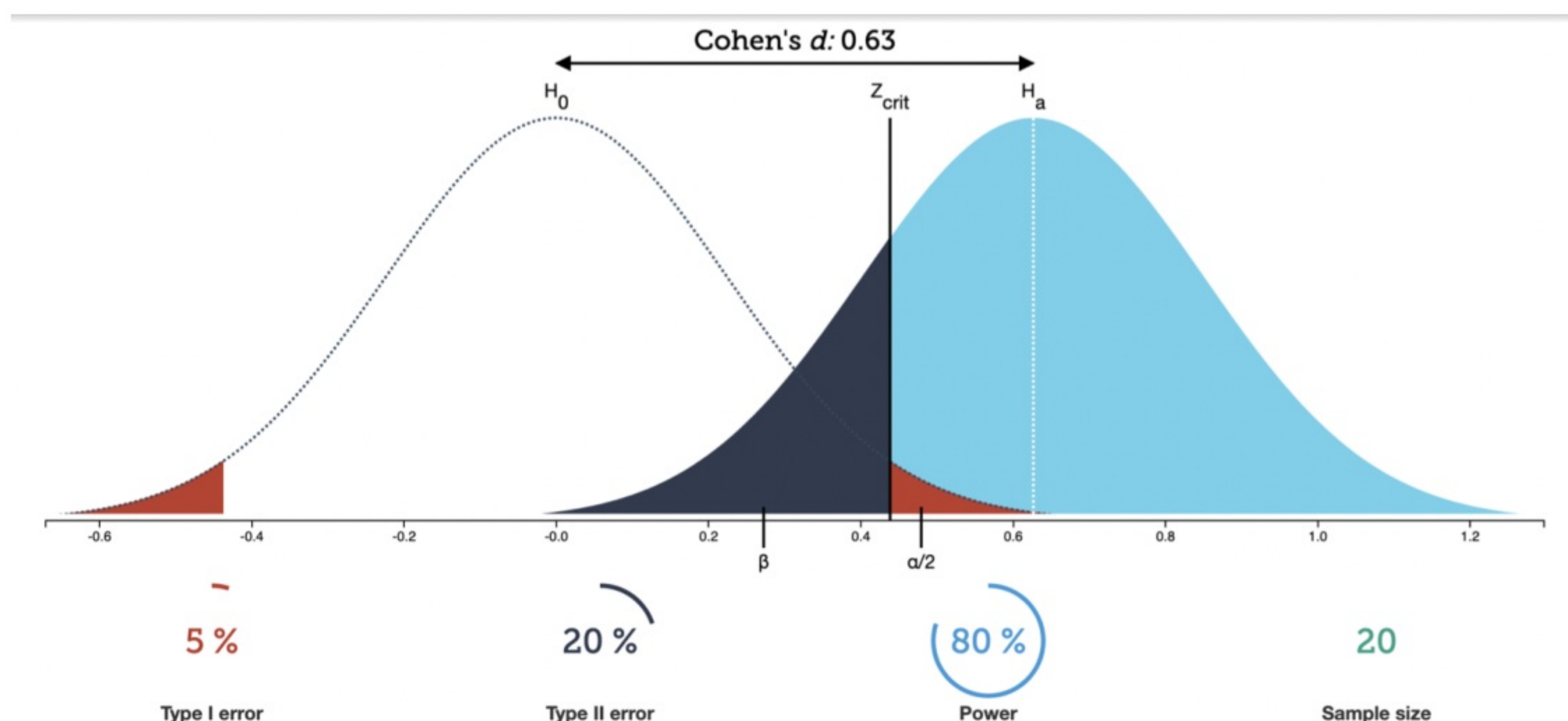
$$\delta(X^n) = \begin{cases} H_0, & \text{если } t(X^n) < C, \\ H_1, & \text{если } t(X^n) \geq C \end{cases}$$

Определение

Уровень значимости — вероятность отклонить нулевую гипотезу при условии её истинности, вероятность совершения ошибки первого рода.

Определение

Статистическая мощность — вероятность отклонения основной гипотезы в случае, когда альтернативная гипотеза верна. Чем выше мощность теста, тем меньше вероятность совершить ошибку второго рода.



> Критическая область

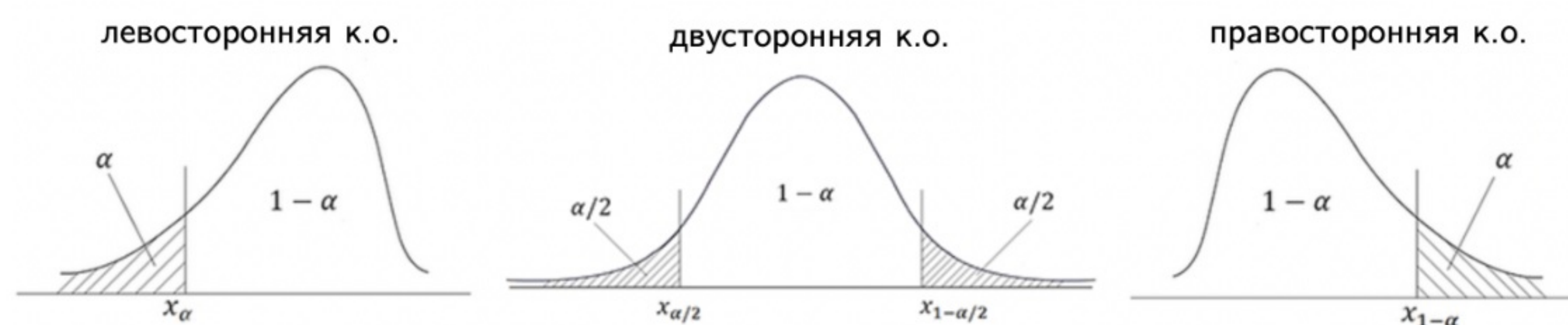
Есть две выборки: $X_1, \dots, X_n \sim F_1$ и $Y_1, \dots, Y_n \sim F_2$

Определим гипотезы: $H_0 : EX = EY$ и $H_1 : EX \neq EY$

Рассмотрим распределение случайной величины: $t = \langle X^n \rangle - \langle Y^n \rangle$.

Определение

Критическая область — область выборочного пространства, при попадании в которую нулевая гипотеза отклоняется.



> Этапы проверки статистических гипотез

1. **Выдвижение** основной **гипотезы** H_0 и альтернативной гипотезы H_1 .
2. **Выбор уровня значимости** α , на котором будет сделан вывод о справедливости гипотезы. Он равен вероятности допустить ошибку первого рода.

3. **Расчёт статистики критерия** такой, что она зависит от выборки и по её значению можно сделать вывод об истинности нулевой гипотезы.
 4. Построение критической области.
 5. По попаданию или не попаданию значения статистики в критическую область делается **вывод об истинности выдвинутой гипотезы** на выбранном уровне значимости.
-

> Важные статистические тесты

> Тест Стьюдента

Есть две выборки: $X_1, \dots, X_{n_1} \sim N(\mu_1, \sigma_1)$ и $Y_1, \dots, Y_{n_2} \sim N(\mu_2, \sigma_2)$.

Гипотезы:

$H_0 : EX = EY$ и $H_1 : EX \neq EY$.

Средние выборок:

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i, \quad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$$

Оценки дисперсий:

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

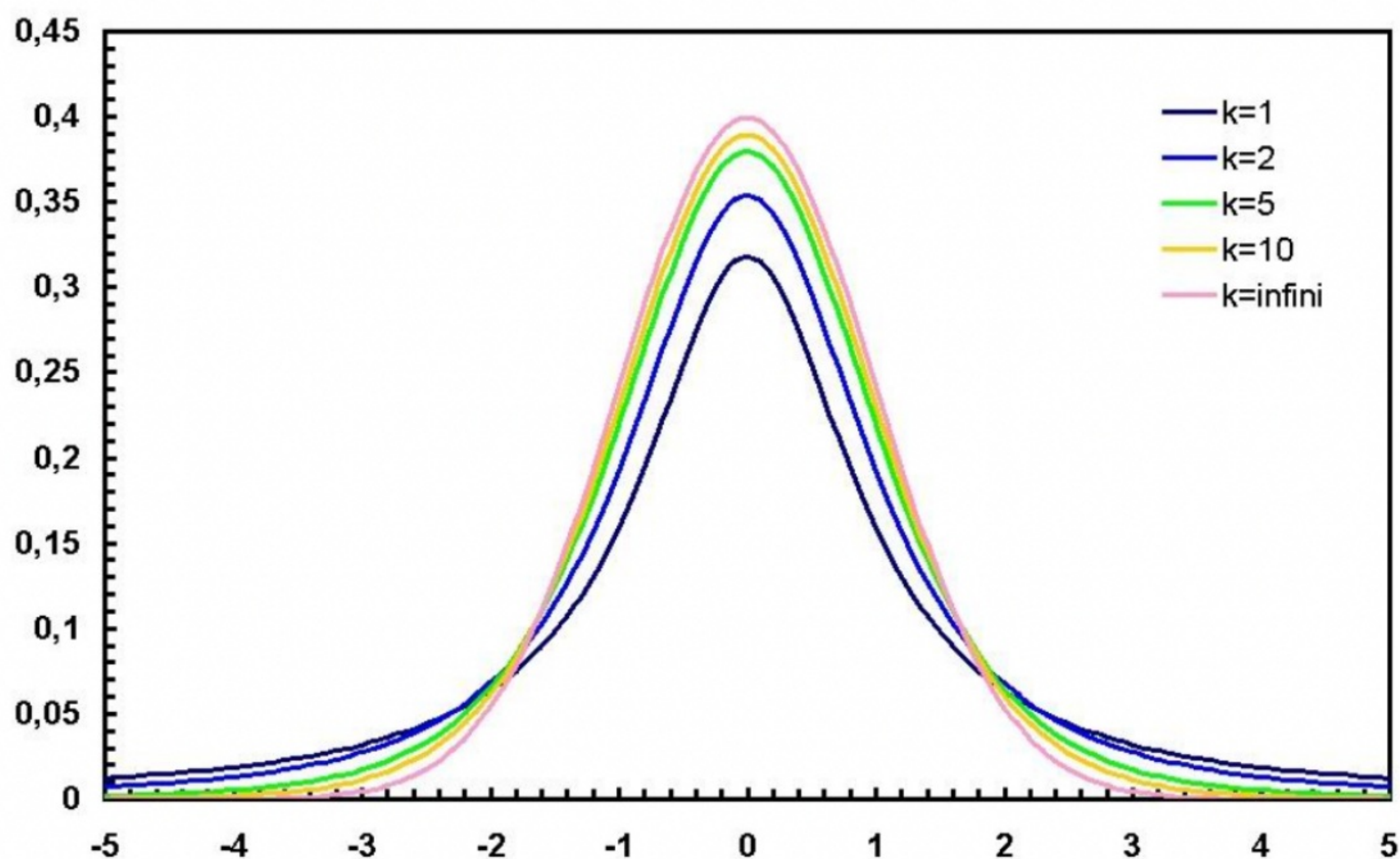
Статистика теста:

$$t(X^n, Y^n) = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \approx St(v)$$
$$v = \frac{(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2})^2}{\frac{S_X^4}{n_1^2(n_1-1)} + \frac{S_Y^4}{n_2^2(n_2-1)}}$$

Предположения:

- **Средние** значения выборок **распределены нормально**;
- Дисперсии выборок **равны**;

- Выборки **независимы** друг от друга.



Плотность распределения Стьюдента

При неизвестных дисперсиях распределение статистики t имеет приближённое распределение (проблема Беренса — Фишера).

Приближение достаточно точное при выполнении одного из следующих условий:

1. $n_1 = n_2$
2. $I(n_1 > n_2) = I(\sigma_1 > \sigma_2)$

> Тест Манна-Уитни

Есть две выборки: $X_1, \dots, X_{n_1} \sim F_X$ и $Y_1, \dots, Y_{n_2} \sim F_Y$.

Гипотезы:

$H_0 : F_X(t) = F_Y(t)$ и $H_1 : F_X(t) = F_Y(t + \Delta), \Delta \neq 0$

1. Составить единый ранжированный ряд из обеих сопоставляемых выборок.
2. Подсчитать отдельно сумму рангов выборок R_1 и R_2 .
3. Вычислить

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2}, U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$$

Значение U-статистики Манна-Уитни:

$$U = \min\{U_1, U_2\}$$

Пример:

Две выборки $X = \{1, 4\}$, $Y = \{2, 5, 7\}$.

Объединим выборки $\{1, 2, 4, 5, 7\}$. Суммы рангов $R_1 = 4$, $R_2 = 11$

$$U_1 = 4 - 3 = 1, U_2 = 11 - 6 = 5$$

$$U = \min\{1, 5\} = 1$$

Свойства:

- $0 \leq U_1 \leq n_1 n_2, 0 \leq U_2 \leq n_1 n_2$
 - $U_1 + U_2 = n_1 n_2$
 - При $n_1, n_2 \geq 20$ распределение $U(X^n, Y^n) \sim N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right)$
-

Предположения:

- X и Y из одного распределения с точностью до сдвига;
 - Элементы внутри выборок независимы;
 - Выборки независимы друг от друга.
-

Замечания:

- Если есть дублирующиеся значения, то для них нужно проставить их средний ранг и внести корректировку в аппроксимирующее нормальное распределение.
 - Устойчив к выбросам, результат может отличаться от теста Стьюдента.
-

> Критерий Колмогорова

Дана выборка $X_1, \dots, X_n \sim F$.

Гипотезы:

$H_0 : F = F_0$ и $H_1 : F \neq F_0$.

Если F_0 непрерывная, то можно пользоваться критерием Колмогорова.

Определим **статистику** $t(X^n) = \sqrt{n} \sup_x |\hat{F}_n(x) - F_0(x)|$.

Если H_0 верна, то статистика $t(X^n)$ имеет распределение Колмогорова:

$$K(x) = \begin{cases} \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 x^2} & , x > 0 \\ 0 & , x \leq 0 \end{cases}$$

Распределение Колмогорова табулировано.

$$\delta(X^n) = \begin{cases} H_0, & \text{если } t(X^n) < C, \\ H_1, & \text{если } t(X^n) \geq C \end{cases}$$

Критерий Колмогорова

> Критерий отношения правдоподобия

Функция правдоподобия $L(\theta) = \prod_{i=1}^n p(X_i; \theta)$.

Гипотезы:

$H_0 : \theta \in \Theta_0$ и $H_1 : \theta \notin \Theta_0$.

Статистика отношения правдоподобия:

$$\lambda(X^n) = 2 \ln \left(\frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} \right) = 2 \ln \left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} \right)$$

где $\hat{\theta}$ — ОМП, $\hat{\theta}_0$ — ОМП при условии $\theta \in \Theta_0$.

Допустим $\theta = (\theta_1, \dots, \theta_q, \theta_{q+1}, \dots, \theta_r)$. Пусть $\Theta_0 = \{\theta : (\theta_{q+1}, \dots, \theta_r) = (\theta_{0,q+1}, \dots, \theta_{0,r})\}$

Если H_0 верна, то

$$\lambda(X^n) \sim \chi^2_{r-q, \alpha}$$

где $r - q$ — размерность Θ минус размерность Θ_0 , α — уровень значимости.

$$\delta(X^n) = \begin{cases} H_0, & \text{если } \lambda(X^n) < C, \\ H_1, & \text{если } \lambda(X^n) \geq C \end{cases}$$

Критерий отношения правдоподобия

> Бутстреп

> Доверительный интервал

Доверительным интервалом с доверительной вероятностью $1 - \alpha$ для параметра θ называется интервал $C_n = (a, b)$, где $a = a(X_1, \dots, X_n)$ и $b = b(X_1, \dots, X_n)$ — такие функции выборки, что $P(\theta \in C_n) \geq 1 - \alpha$

Возьмём в качестве параметра θ разность средних распределений. Нулевая гипотеза:

$$H_0 : EX = EY$$

Тогда критерий проверки гипотезы о равенстве средних будет иметь вид:

$$0 \notin (a, b) \Leftrightarrow \text{отвергнуть гипотезу } H_0$$

Примеры расположения доверительного интервала **относительно нуля**:

В первом случае **значимых отличий нет, ноль внутри** доверительного интервала.

Во втором случае **значимые отличия есть, ноль вне** доверительного интервала.



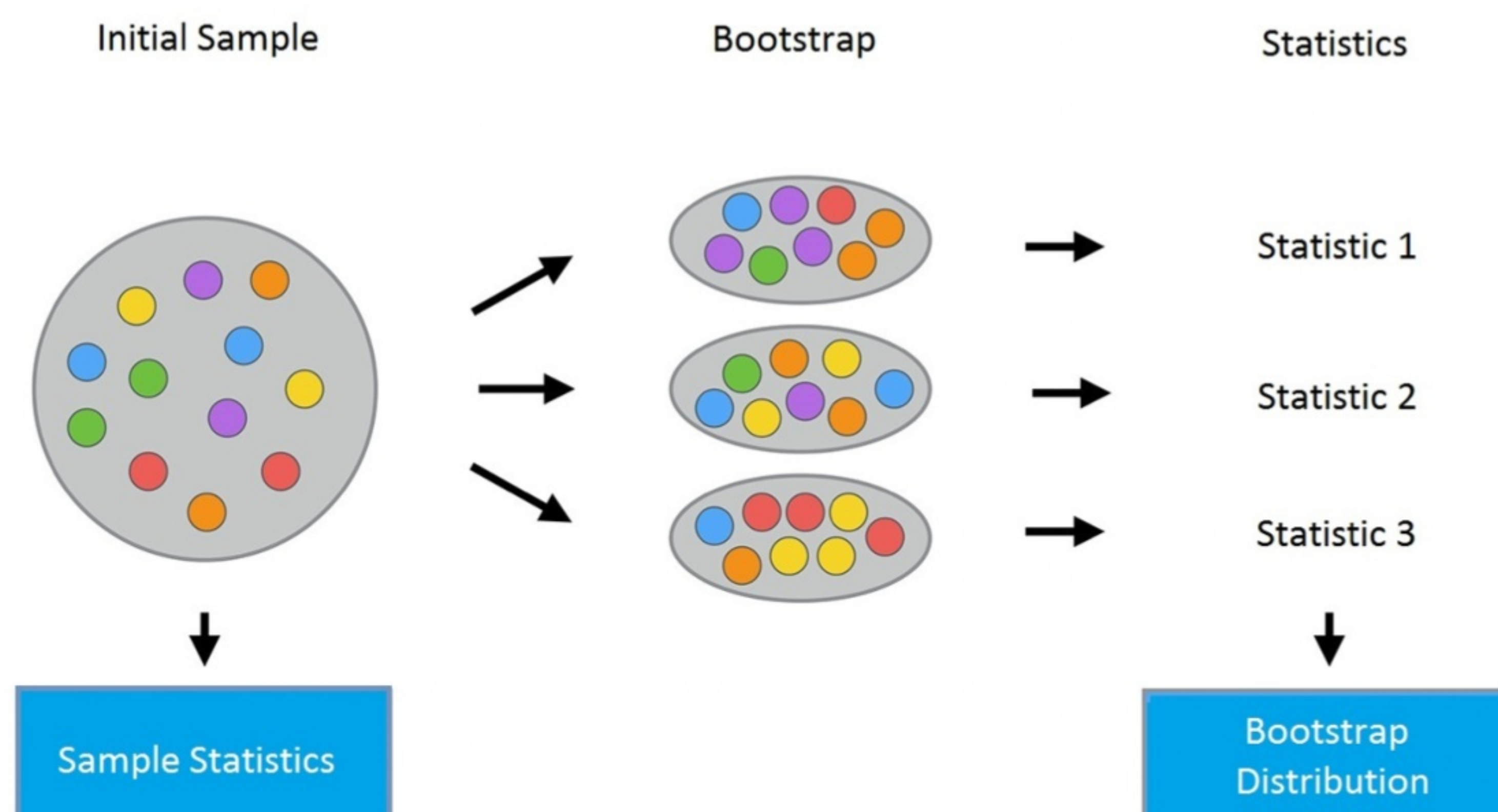
> Бутстреп

Определение

Бутстреп — это метод для подсчёта стандартных ошибок и нахождения доверительных интервалов статистических функционалов.

Хотим оценить распределение статистики T по выборке $X_1, \dots, X_n$.

1. Генерируем B подвыборок из выборки X_n .
2. Вычисляем статистику T для каждой подвыборки.
3. Оцениваем распределение по получившемуся множеству статистик.



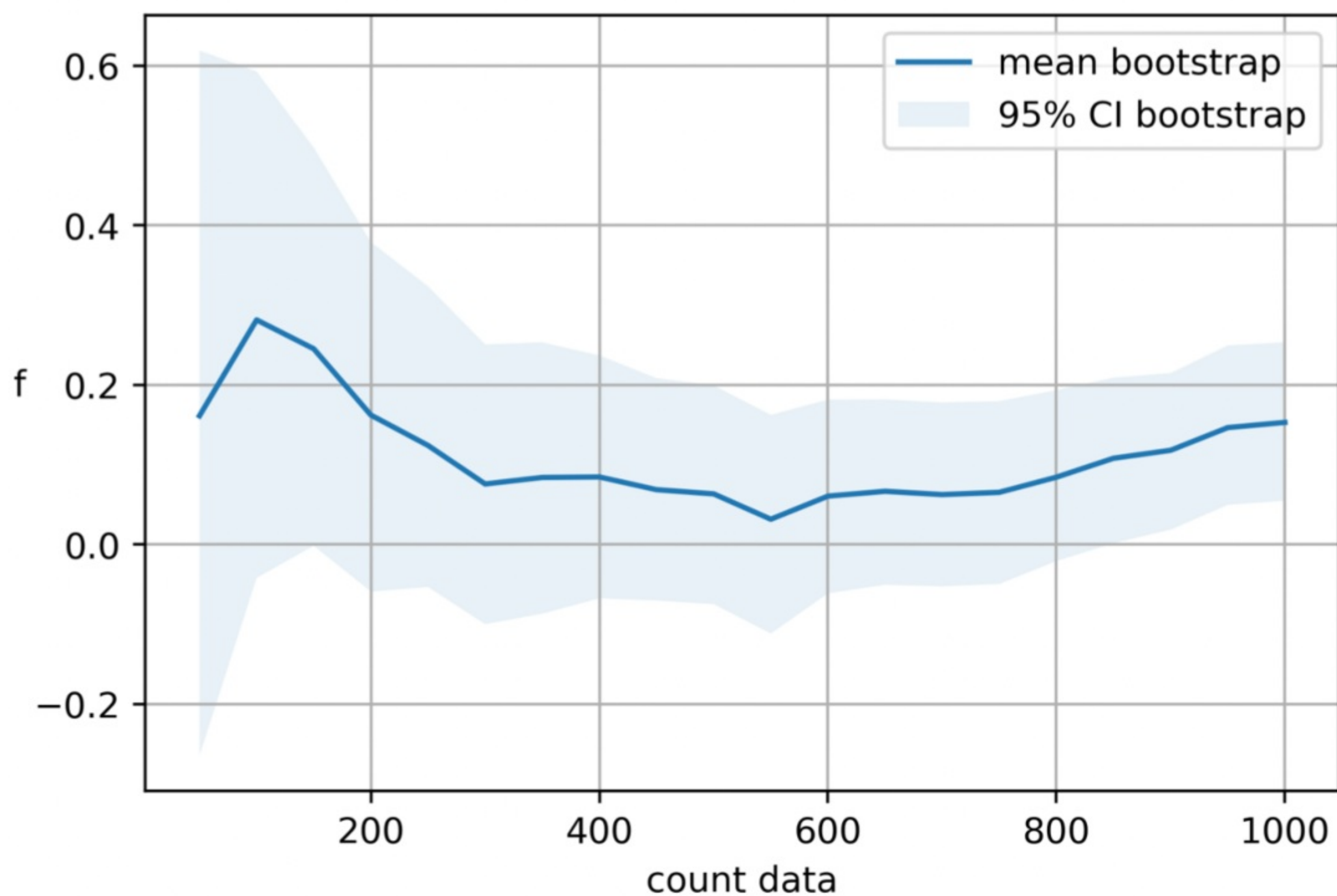
Пример

Дана функция:

$$f(x) = x \cos(71x) + \frac{\sin(13x^2)}{x}$$

Хотим оценить $m = E(f(X))$ по выборке $X_1, \dots, X_n \sim N(1, 1)$. Точечная оценка

$$\hat{m} = n^{-1} \sum_i f(X_i)$$



Для оценки разброса оценки воспользуемся бутстрепом

> Проверка гипотезы о равенстве средних

Есть две выборки: $X_1, \dots, X_n \sim F_X$ и $Y_1, \dots, Y_n \sim F_Y$.

Гипотезы:

$H_0 : EX = EY$ и $H_1 : EX \neq EY$.

Генерируем B пар подвыборок из выборок X_n, Y_n размером n .

Для каждой пары считаем разность выборочных средних $T_{n,1}, \dots, T_{n,B}$.

По получившимся множеству разностей строим доверительный интервал и проверяем гипотезу.

> Нормальный интервал

Предположим, что полученное множество статистик, посчитанных на бутстрепных данных, имеет **нормальное распределение**. Тогда

$$C_n = (T - z_{\alpha/2} \hat{se}_{boot}, T + z_{\alpha/2} \hat{se}_{boot}),$$

где **разность выборочных средних** $T = \overline{Y^n} - \overline{X^n}$, **оценка стандартной ошибки** на основе бутстрепа:

$$\hat{se}_{boot} = \sqrt{1/B \sum_{i=1}^B (T_{n,i} - \overline{T_n^B})^2}, \text{ модуль квантиля стандартного нормального распределения } z_{\alpha/2}.$$

> Интервал на основе процентилей

$$C_n = (\theta_{\alpha/2}^*, \theta_{1-\alpha/2}^*),$$

где θ_{α}^* — квантили, посчитанные по множеству $\{T_{n,1}, \dots, T_{n,B}\}$.

Можно подобрать монотонное преобразование, которое преобразует распределение статистики T в распределение, похожее на нормальное. Так как монотонное преобразование сохраняет квантили, то квантили $\theta_{\alpha/2}$, $\theta_{1-\alpha/2}$ перейдут в соответствующие квантили нормального распределения.

Тогда легко показать, что вероятность попасть в определённый выше интервал равна $1 - \alpha$.

> Центральный интервал

Пусть $\theta = T(F)$ и $\hat{\theta}_n = T(\hat{F}_n)$.

Введём $R_n = \hat{\theta}_n - \theta$ с распределением $H(r) = P_F(R_n \leq r)$.

Определим **доверительный интервал**:

$$C_n = (a, b) = (\hat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2}), \hat{\theta}_n - H^{-1}(\frac{\alpha}{2}))$$

Легко показать, что $P(\theta \in C_n) = 1 - \alpha$, но мы не знаем $H(r)$.

Оценим его с помощью бутстрепа:

$$\hat{H}(r) = \frac{1}{B} \sum_{i=1}^B I(R_{n,i}^* \leq r), \quad R_{n,i}^* = \hat{\theta}_{n,i}^* - \hat{\theta}_n$$

Пусть r_{β}^* - β -выборочный квантиль, посчитанный по выборке $(R_{n,1}^*, \dots, R_{n,B}^*)$, а θ_{β}^* β -выборочный квантиль, посчитанный по выборке $(\hat{\theta}_{n,1}^*, \dots, \hat{\theta}_{n,B}^*)$, тогда

$$\hat{a} = \hat{\theta} - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r_{1-\alpha/2}^* = 2\hat{\theta}_n - \theta_{1-\alpha/2}^*$$

$$\hat{b} = \hat{\theta} - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r_{\alpha/2}^* = 2\hat{\theta}_n - \theta_{\alpha/2}^*$$

Получаем центральный доверительный интервал:

$$C_n = (2\hat{\theta}_n - \theta_{1-\alpha/2}^*; 2\hat{\theta}_n - \theta_{\alpha/2}^*)$$

Заметим, что $P(T(F) \in C_n) \rightarrow 1 - \alpha$ при $n \rightarrow \infty$.

Таким образом, бутстреп:

- Позволяет оценить распределение некоторой функции от случайной выборки;
- Не делает предположений о виде распределения;
- Требуется много вычислений.

$$C_n = (a, b) = (\hat{\theta}_n - H^{-1}(1 - \frac{\alpha}{2}), \hat{\theta}_n - H^{-1}(\frac{\alpha}{2}))$$

> Как выбрать критерий

> Ошибки I и II рода



Нашли эффект, когда его нет

(ошибка I рода)

“Новые стеллажи увеличат
средний чек на 1%.”

Установить 14000 стеллажей,
которые ничего не меняют.



Не нашли эффект, когда он был

(ошибка II рода)

Смена ассортимента увеличивает
выручку на 2%

Провели пилот на 5 магазинах, тк
дорого оборудовать.
Не увидели стат значимого эффекта.



> Оценка ошибки I рода

1. Генерируем пилотную и контрольную группы.
2. На исторических данных, где не был запущен эксперимент, считаем метрики для групп.
3. Оцениваем значимость отличия средних и запоминаем результат.
4. Повторяем первые три пункта, чтобы набрать статистику.
5. Считаем долю случаев, когда средние значения отличались значимо.

> Оценка ошибки II рода

1. Генерируем пилотную и контрольную группы.
2. На исторических данных, где не был запущен эксперимент, считаем метрики для групп, к метрикам пилотной группы добавляем эффект.
3. Оцениваем значимость отличия средних и запоминаем результат.
4. Повторяем первые три пункта, чтобы набрать статистику.
5. Считаем долю случаев, когда средние значения не отличались значимо.

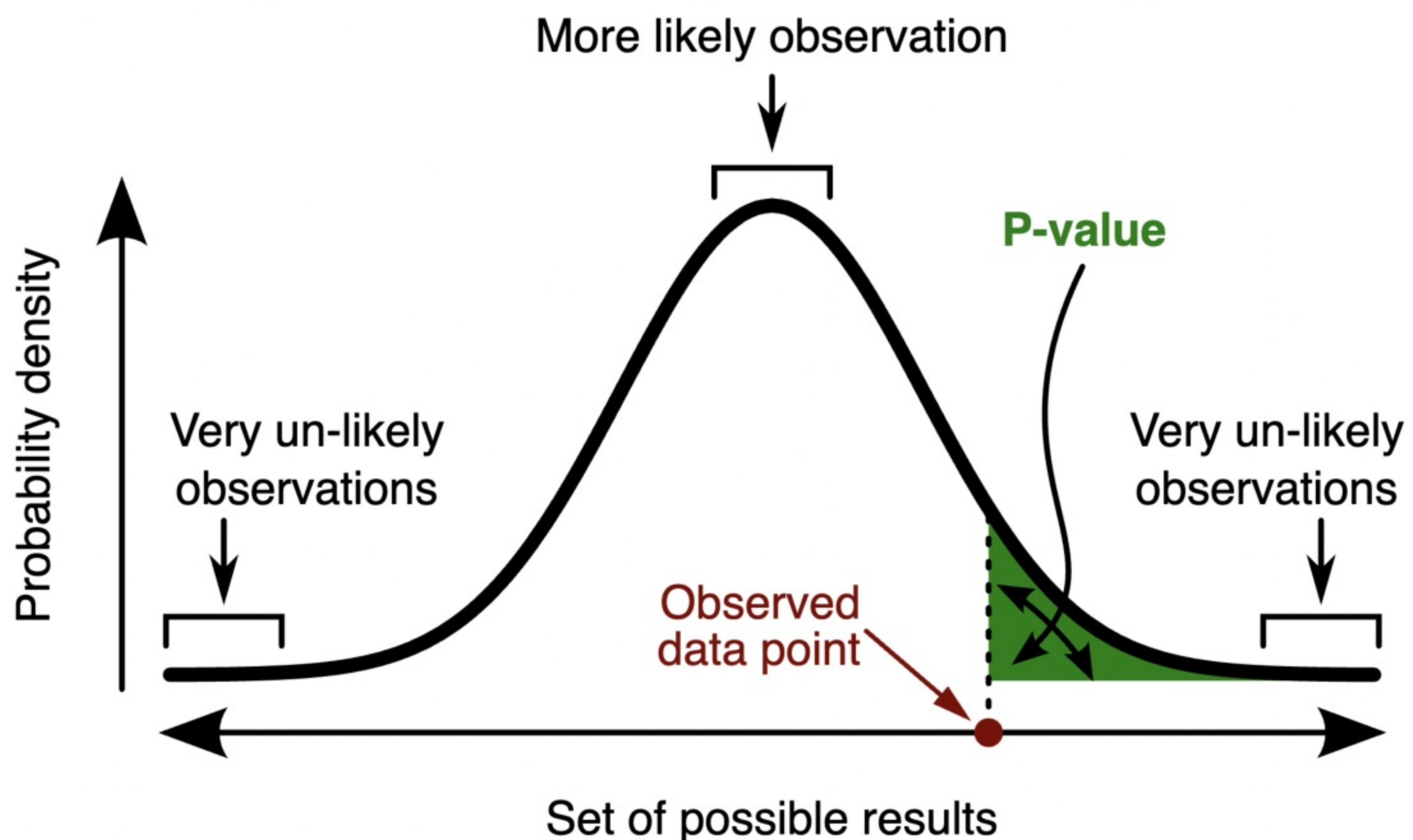
> p-value

p_{value} — вероятность при нулевой гипотезе наблюдать полученное или более экстремальное значение статистики.

$$p_{value} = P(T > t | H_0)$$

Статистический критерий можно записать как:

$$p_{value} < \alpha \Leftrightarrow \text{отвергнуть гипотезу.}$$



> Распределение p-value

Теорема

Пусть тест размера α имеет вид: отвергнуть $H_0 \Leftrightarrow T(x^n) > c(\alpha)$, где x^n — наблюдаемая выборка.

Если H_0 верна, то $p_{value}(x^n) = P(T(X^n) > T(x^n) | H_0)$.

Из последнего свойства также следует, что $p_{value}(X^n) \sim Uniform(0, 1)$.

Утверждение

Пусть случайная величина X имеет распределение $F(x)$, и $F(x)$ обратима. Тогда случайная величина $Y = F(X)$ имеет распределение $Uniform(0, 1)$.

Доказательство:

$$P(F(X) < x) = P(X < F^{-1}(x)) = F(F^{-1}(x)) = x, x \in (0, 1).$$

Случайная величина $Y = 1 - F(X)$ также является равномерной, причём:

$$Y = 1 - F(X) = P_{Z \sim F(\cdot)}(Z > X),$$

Теперь p_{value} подходит под роль Y в утверждении выше: $Y = p_{value}(X^n), X = T(X^n)$.
