Проверка статистических гипотез

Теория. Типичные ошибки.

Антон Коробейников

Летняя школа по биоинформатике

27 июля 2016 года

Содержание

Проверка статистических гипотез

Зачем:

- Основа для статистического вывода

Проверка статистических гипотез

Зачем:

- Основа для статистического вывода

Легко сделать ошибку:

- Непривычная логика
- Рецепты: применение критерия в неправильных предположениях
- Неверная интерпретация результатов
- Игнорирование эффектов зависимости от объема выборки, множественных сравнений и пр.

Постановка задачи

Выборка: $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathsf{P} \in \mathcal{P}$

Гипотеза: H_0 — некоторое высказывание относительно распределения $\mathbf{X}$. $\mathsf{P}_0 \in \mathcal{P}$.

Постановка задачи

Выборка: $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathsf{P} \in \mathcal{P}$

Гипотеза: H_0 — некоторое высказывание относительно распределения $\mathbf{X}$. $\mathsf{P}_0 \in \mathcal{P}$.

Задача: По выборке $\mathbf{X}$ принять или отвергнуть H_0.

Постановка задачи

Выборка: $\mathbf{X} = (X_1, \ldots, X_n) \sim \mathsf{P} \in \mathcal{P}$

Гипотеза: H_0 — некоторое высказывание относительно распределения $\mathbf{X}$. $\mathsf{P}_0 \in \mathcal{P}$.

Задача: По выборке $\mathbf{X}$ принять или отвергнуть H_0.

Задача*: Предъявить алгоритм, решающий задачу.

Примеры

- $\mathcal{P}$ — семейство нормальных распределений $N(a, 1)$, $a \in \mathbb{R}$.
  $H_0 : a = 0$.
- $\mathcal{P}$ — семейство всех распределений на прямой.
  $H_0 : \mathbf{E}\xi = 0$.
- $\mathcal{P}$ — семейство всех распределений на прямой.
  $H_0 : \mathsf{P}_0 \in \left\{N(a, \sigma^2\right\}$.

Критерий

Оракул:

$$\delta(\mathbf{X}) = \delta(X_1, \ldots, X_n) = \begin{cases} 0, & \text{принять } H_0, \\ 1, & \text{отвергнуть } H_0. \end{cases}$$

Критическая область:

$$S = \{\mathbf{X} : \delta(\mathbf{X}) = 1\}$$

Доверительная область:

$$D = \{\mathbf{X} : \delta(\mathbf{X}) = 0\}$$

Ошибки

- Критерий может (и должен!) ошибаться
- Вы должны это постоянно учитывать
- При многоступенчатых процедурах ошибки могут накапливаться и умножаться

Ошибки

Ошибка первого рода: Отвергнуть H_0 тогда, когда она верна.

Ошибка второго рода: Не отвергать H_0, когда она не верна.

Ошибки

Ошибка первого рода: Отвергнуть H_0 тогда, когда она верна.

Ошибка второго рода: Не отвергать H_0, когда она не верна.

В терминах критической и доверительной области критерия:

Ошибка первого рода: H_0 верна, $\mathbf{X} \in S$.

Ошибка второго рода: H_0 не верна, $\mathbf{X} \in D$

Как строить критерии?

- Как выбрать разумно $\delta(\mathbf{X})$?
- Как сравнивать разные критерии? По каким характеристикам?

Вероятности ошибок первого и второго рода

Вероятность ошибки первого рода: $\alpha_I = \mathsf{P}(\delta(\mathbf{X}) \in S | H_0)$

Вероятность ошибки второго рода: $\alpha_{II} = \mathsf{P}(\delta(\mathbf{X}) \notin S | H_1)$

Ошибки не симметричны!

- $\alpha_I = \mathsf{P}(\delta(\mathbf{X}) \in S | H_0) = \mathsf{P}_0(\delta(\mathbf{X}) \in S).$
- $\alpha_{II} = \mathsf{P}(\delta(\mathbf{X}) \notin S | H_1) = \mathsf{P}_?(\delta(\mathbf{X}) \notin S)$

Простейший пример критерия

$$\mathcal{P} = \{N(a_0, 0.42), N(a_1, 0.42)\}$$

$$H_0 : N(a_0, 0.42) \qquad H_1 : N(a_1, 0.42)$$

Статистика критерия:

$$t(\mathbf{X}) = \sqrt{n}\frac{\bar{X} - a_0}{\sqrt{0.42}}$$

Критерий:

$$\delta(\mathbf{X}) = \begin{cases} 0, & |t(\mathbf{X})| < 1.96, \\ 1, & |t(\mathbf{X})| \geq 1.96. \end{cases}$$

Hands on

$$H_0 : N(a_0, 0.42) \qquad H_1 : N(a_1, 0.42)$$

Статистика критерия:

$$t(\mathbf{X}) = \sqrt{n}\frac{\bar{X} - a_0}{\sqrt{0.42}}$$

Критерий:

$$\delta(\mathbf{X}) = \begin{cases} 0, & |t(\mathbf{X})| < 1.96, \\ 1, & |t(\mathbf{X})| \geq 1.96. \end{cases}$$

Оценить:

- Вероятность ошибки первого рода α_I при $n = 100$.
- Вероятность ошибки второго рода α_{II} при $n = 10, 100, 1000$.

Свойства критерия

$$t(\mathbf{X}) = \sqrt{n}\frac{\bar{X} - a_0}{\sqrt{0.42}} \sim_{H_0} N(0,1)$$

Вероятность ошибки первого рода:

$$\alpha_I = \mathsf{P}_0(\delta(\mathbf{X}) = 1) = \mathsf{P}_0(t(\mathbf{X}) \geq 1.96) =$$
$$= \Phi(-1.96) + 1 - \Phi(1.96) = 2\Phi(1.96) - 1 = 0.05$$

Свойства критерия

$$t(\mathbf{X}) = \sqrt{n}\frac{\bar{X} - a_0}{\sqrt{0.42}} \sim_{H_0} N(0,1)$$

Вероятность ошибки первого рода:

$$\alpha_I = \mathsf{P}_0(\delta(\mathbf{X}) = 1) = \mathsf{P}_0(t(\mathbf{X}) \geq 1.96) =$$
$$= \Phi(-1.96) + 1 - \Phi(1.96) = 2\Phi(1.96) - 1 = 0.05$$

Более того, можно построить критическую область для любого *наперед заданного значения* α_I:

$$\alpha = \mathsf{P}_0(t(\mathbf{X}) \geq x_\alpha) = 2\Phi(x_\alpha) - 1,$$
$$x_\alpha = \Phi^{-1}\left(\frac{1+\alpha}{2}\right),$$
$$S_\alpha = (-\infty, -x_\alpha) \cup (x_\alpha, \infty).$$

Свойства критерия

Распределение при H_1:

$$t(\mathbf{X}) = \sqrt{n}\,\frac{\bar{X} - a_0}{\sqrt{0.42}} \sim_{H_1} N\left(\sqrt{n}\,\frac{a_1 - a_0}{\sqrt{0.42}}, 1\right)$$

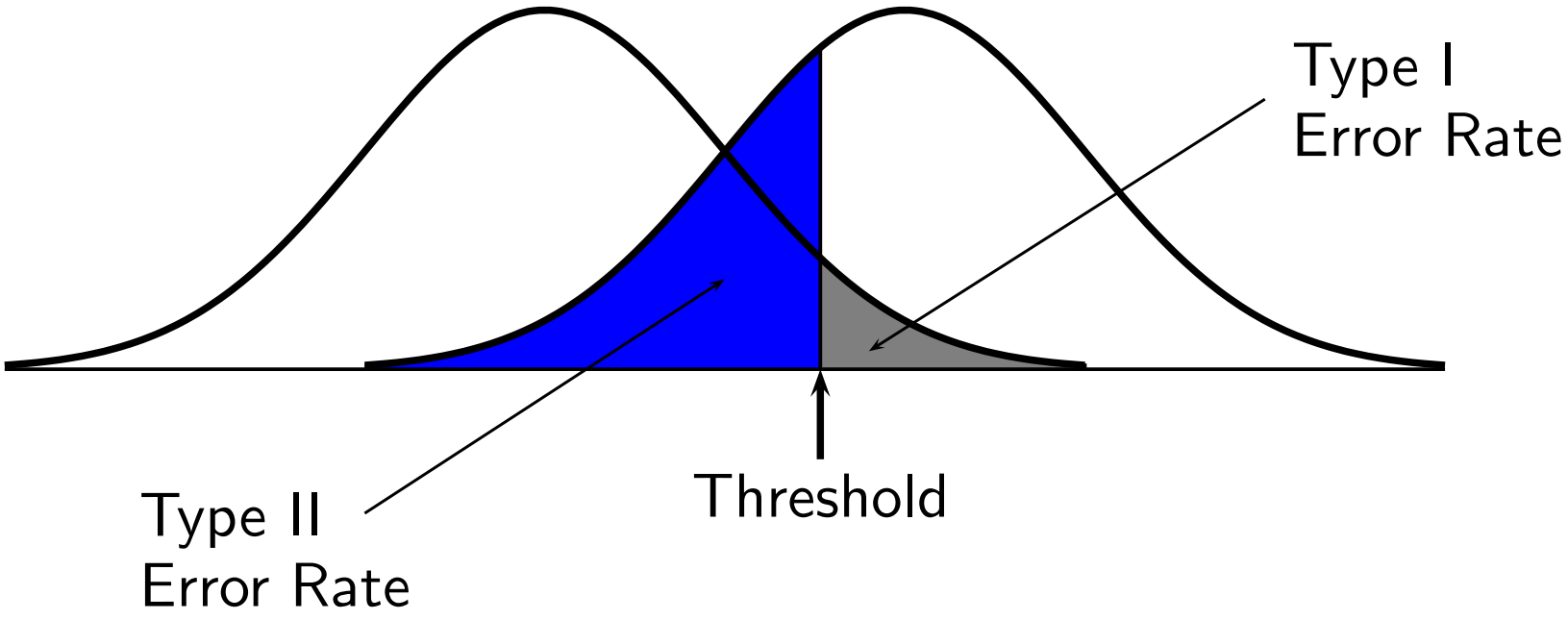
Вероятность ошибки второго рода:

$$\alpha_{II} = \mathsf{P}_1(\delta(\mathbf{X}) = 1) = \mathsf{P}_1(t(\mathbf{X})) =$$
$$= \Phi(x_\alpha - \sqrt{n}\,\frac{a_1 - a_0}{\sqrt{0.42}}) - \Phi(-x_\alpha - \sqrt{n}\,\frac{a_1 - a_0}{\sqrt{0.42}}).$$

Критерий асимптотической мощности 1:

$$\beta = 1 - \alpha_{II} = \mathsf{P}(H_1|H_1) \to 1, \quad n \to \infty.$$

Об ошибках



Общий подход к построению критерия

- Строится $t(\mathbf{X})$ — *статистика критерия*. распределение которой полностью известно в случае выполнения H_0.
- Фиксируется *уровень значимости* α — вероятность ошибки первого рода.
- Строится критическая область $S_\alpha = \mathsf{P}_0\left(t(\mathbf{X}) \in S_\alpha\right) = \alpha$.
- По возможности выбирается S_α так, чтобы минимизировать α_{II} (максимизировать мощность β).
- Гипотеза H_0 отвергается, если $t(\mathbf{X}) \in S_\alpha$.

О p-значениях

- Часто критическую область не строят явно (а зря!)
- p-значение — вероятность наблюдать такое же, или «большее» значение статистики критерия, если верна H_0.
- Формально: $p(\mathbf{X}) = \inf_\alpha \{t(\mathbf{X}) \in S_\alpha\}$
- $\mathsf{P}_0\left(p(\mathbf{X}) < u\right) = \mathsf{P}_0\left(t(\mathbf{X}) \in S_u\right) = u$

О p-значениях

- Часто критическую область не строят явно (а зря!)
- p-значение — вероятность наблюдать такое же, или «большее» значение статистики критерия, если верна H_0.
- Формально: $p(\mathbf{X}) = \inf_\alpha \{t(\mathbf{X}) \in S_\alpha\}$
- $\mathsf{P}_0\left(p(\mathbf{X}) < u\right) = \mathsf{P}_0\left(t(\mathbf{X}) \in S_u\right) = u$

Когда верна H_0 p-значения имеют равномерное на $[0, 1]$ распределение.

Типичные заблуждения:

- p-значение — это не вероятность того, что H_0 верна.
- p значение не показывает, насколько хорошо данные подходят под H_0.

Hands-on: p-значения

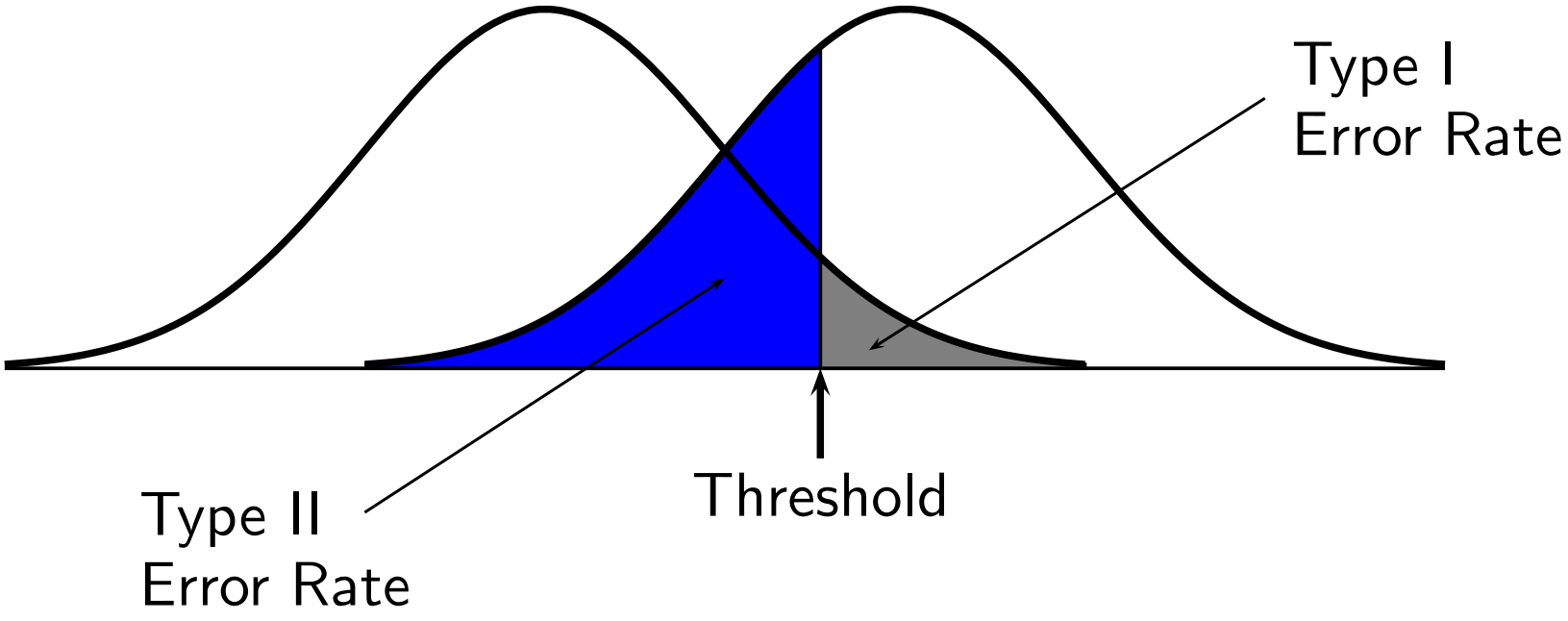
$$H_0 : N(a_0, 0.42) \qquad H_1 : N(a_1, 0.42)$$

Статистика критерия:

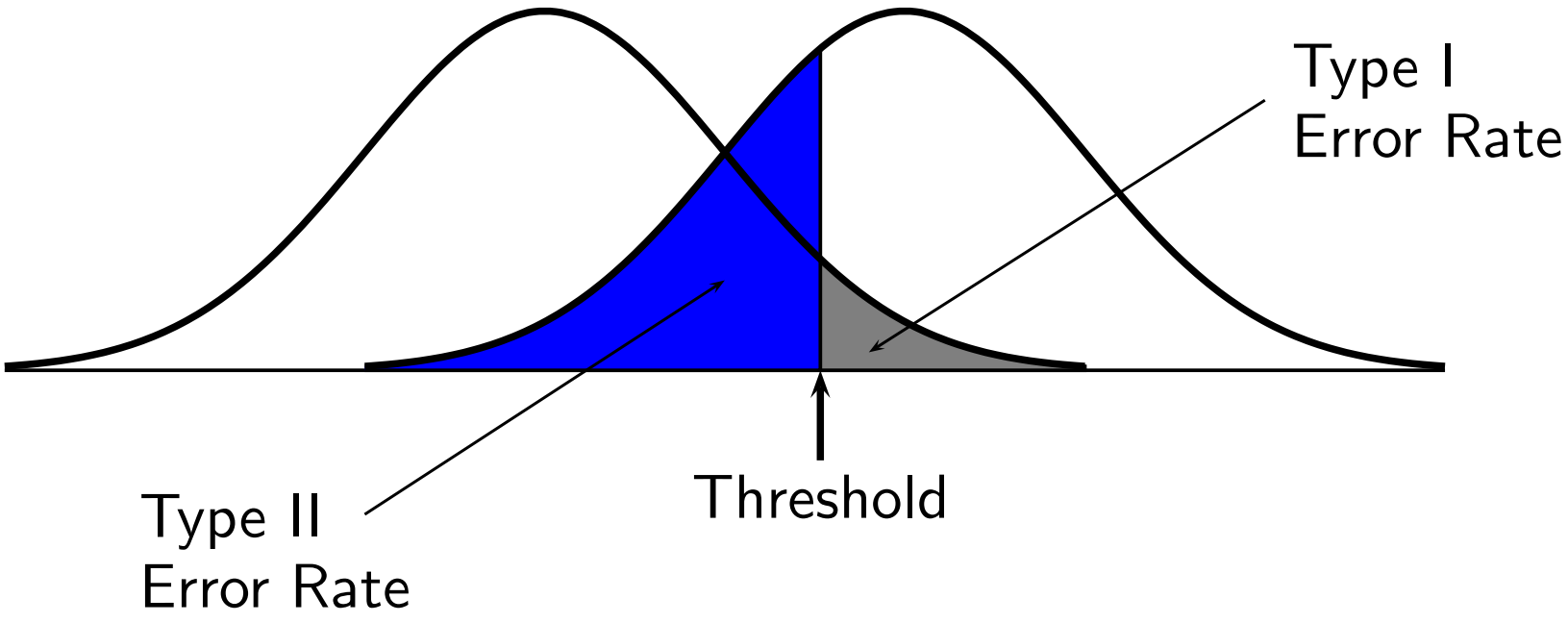
$$t(\mathbf{X}) = \sqrt{n}\frac{\bar{X} - a_0}{\sqrt{0.42}}$$

- Построить распределение для p-значения в случае, когда верна H_0
- Построить распределение для p-значения в случае, когда верна H_1

Еще раз о сравнении критериев



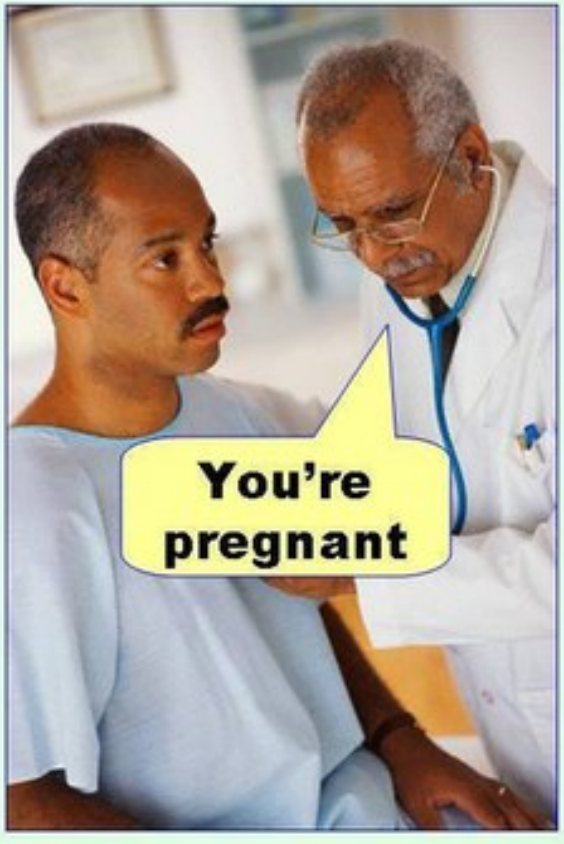
Еще раз о сравнении критериев



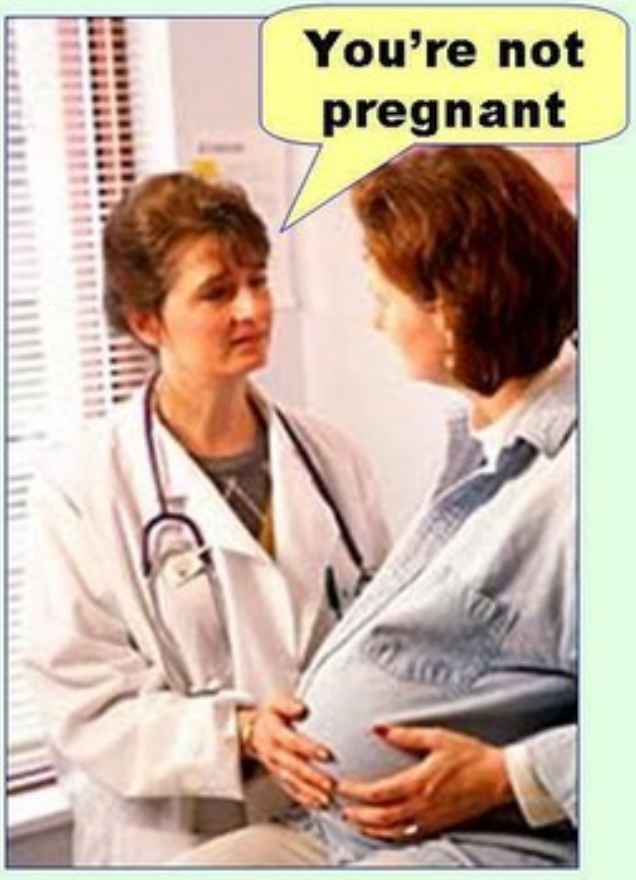
Борьба за мощность: найти тест, обладающей наибольшей мощностью (наименьшей ошибкой второго рода) при фиксированной ошибке первого рода

Еще раз коротко об ошибках

Type I error
(false positive)



Type II error
(false negative)



Одновыборочный t-критерий

Условия:

$$X_1, \ldots, X_n \sim N(a, \theta)$$

Гипотеза:

$$H_0 : a = a_0 \qquad H_1 : a \neq a_0$$

Статистика критерия:

$$t = \sqrt{n} \frac{\bar{X} - a_0}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}} \sim_{H_0} t_{n-1}$$

B R

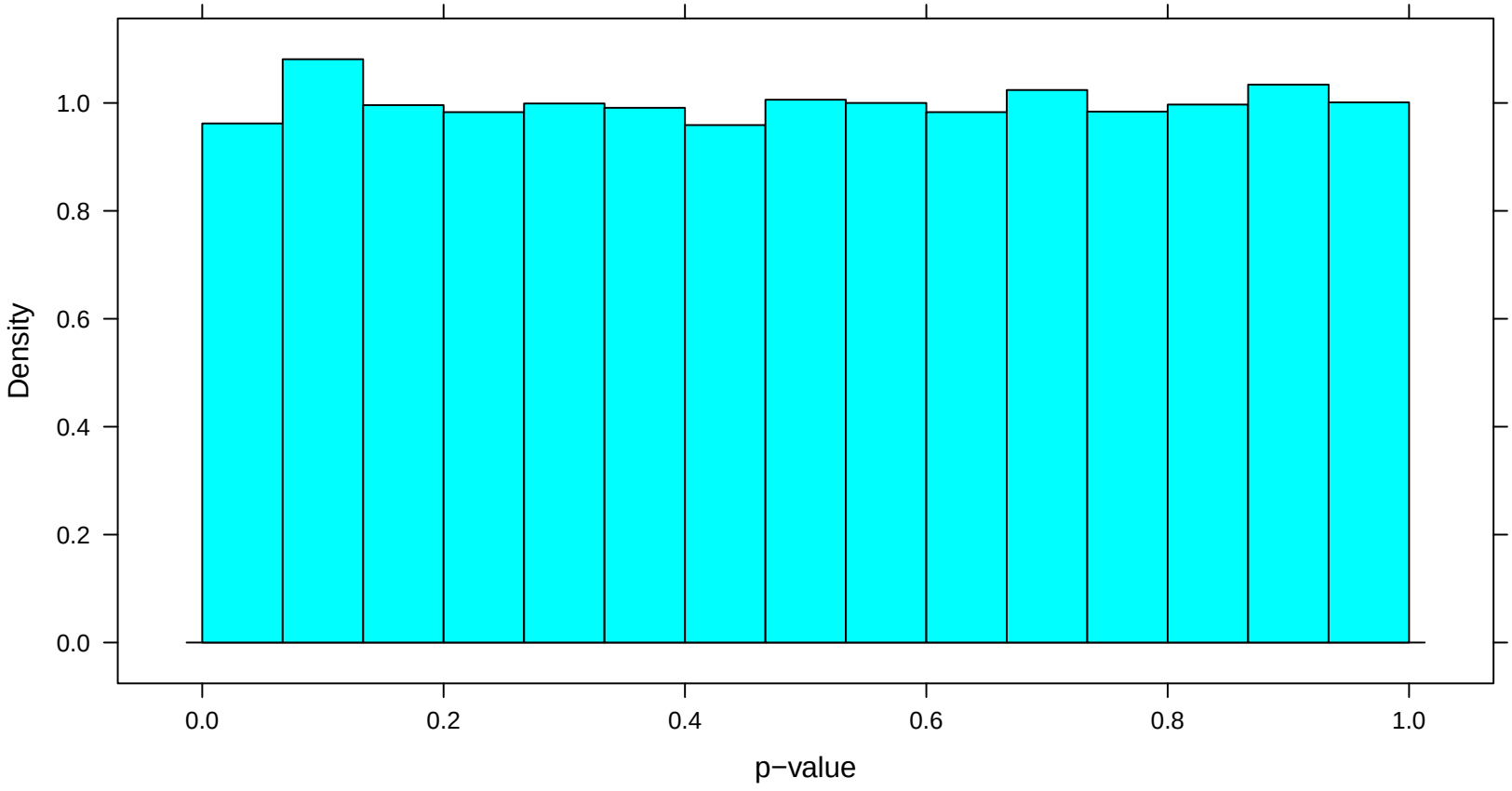
```
> v <- rnorm(10)
> t.test(v, mu = 0)

One Sample t-test

data:  v
t = 1.5291, df = 9, p-value = 0.1606
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1593860  0.8243201
sample estimates:
mean of x
0.3324671
```

B R

```
> v <- rnorm(10)
> tt <- t.test(v, mu = 0.5)
> names(tt)
[1] "statistic"   "parameter"   "p.value"     "conf.int"
[6] "null.value"  "alternative" "method"      "data.name"
> tt$p.value
[1] 0.4607349
```

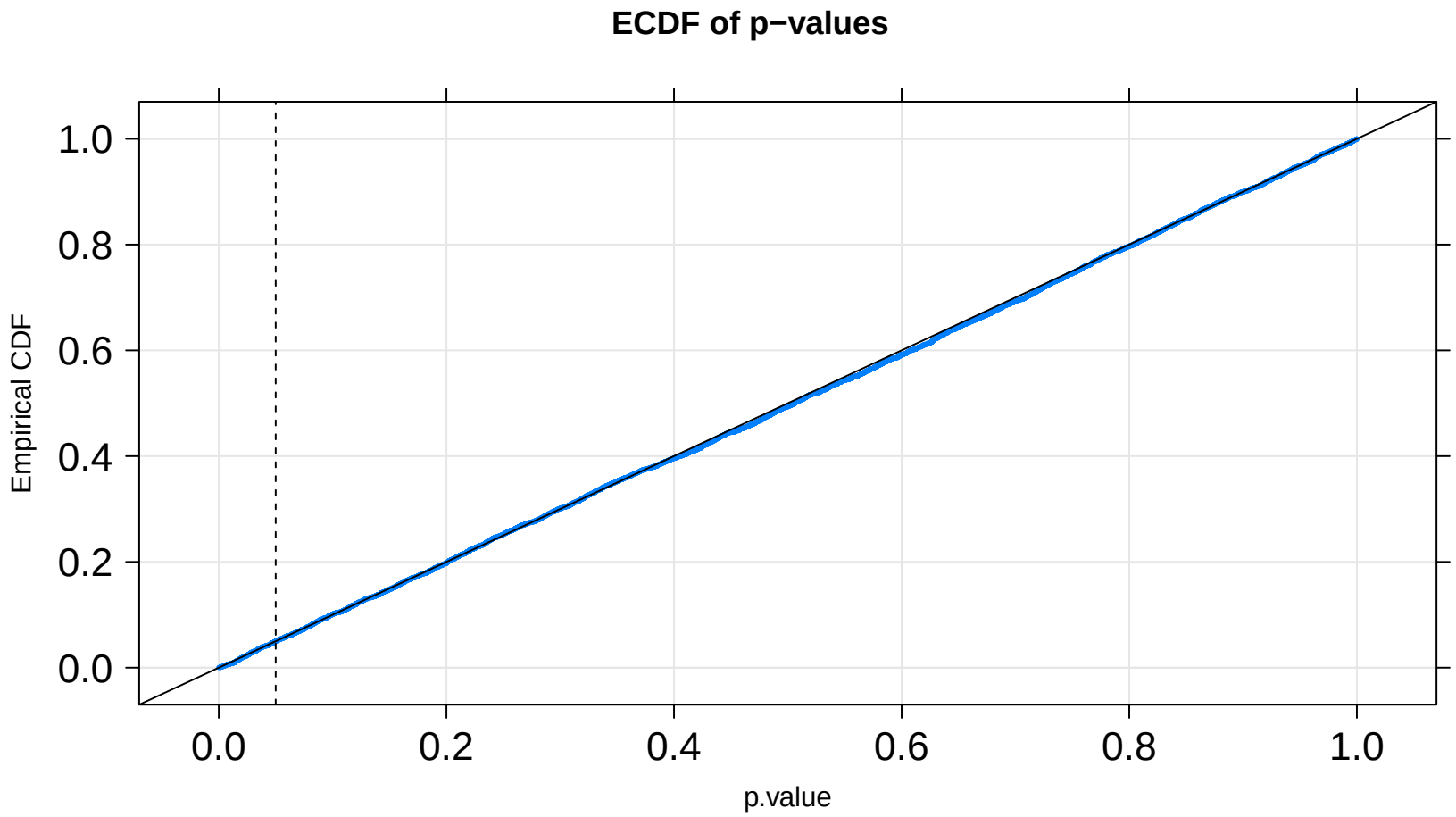
Напоминание про p-значение

Если верна H_0, то p-значения имеют равномерное на отрезке $[0, 1]$ распределение.



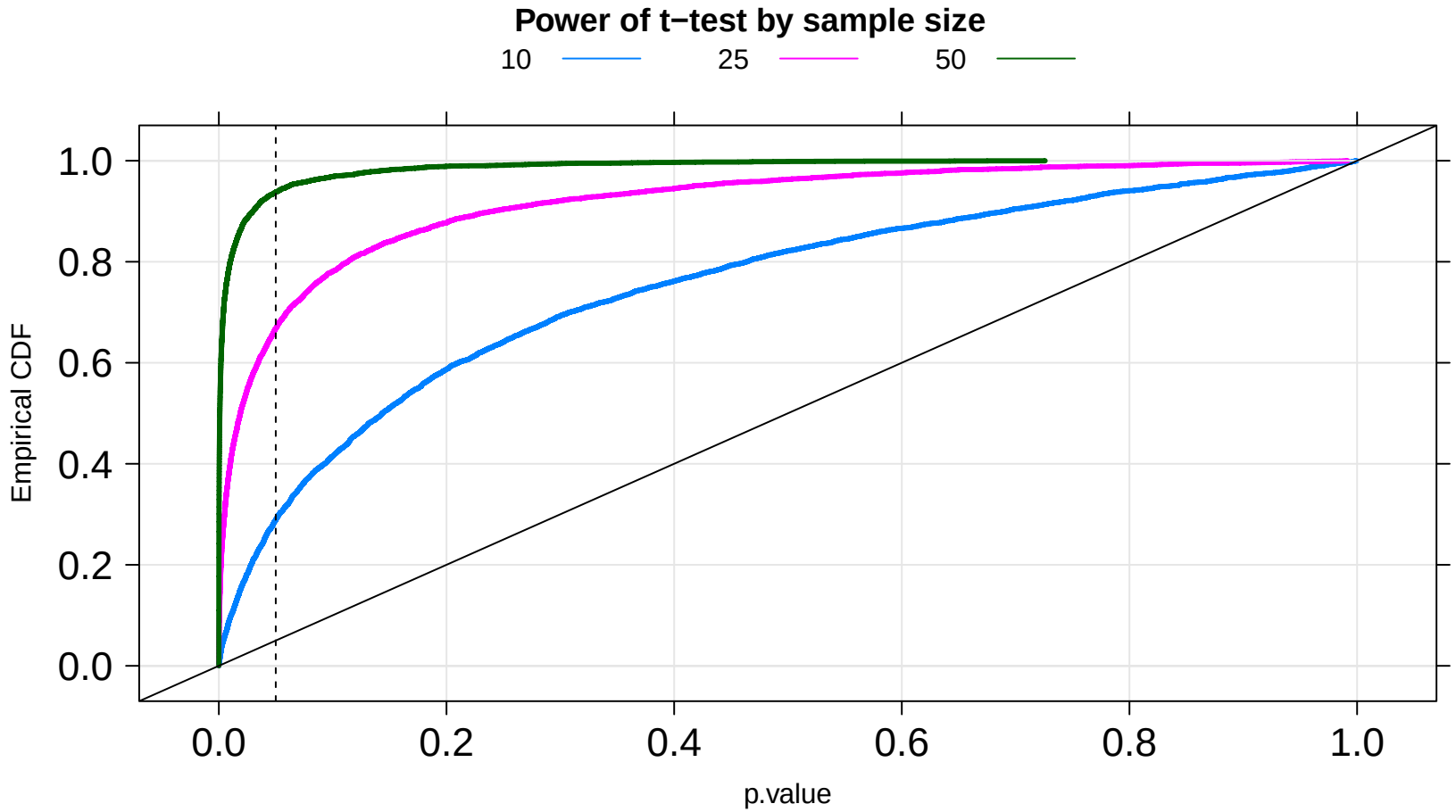
Напоминание про p-значение

Если верна H_0, то p-значения имеют равномерное на отрезке $[0, 1]$ распределение.



Напоминание про p-значение

Иначе (если H_0 не верна) мы получаем график зависимости мощности от вероятности ошибки первого рода.



Hands-on

Условия:

$$X_1, \ldots, X_n \sim N(a, \theta)$$

Гипотеза:

$$H_0 : a = a_0 \qquad H_1 : a \neq a_0$$

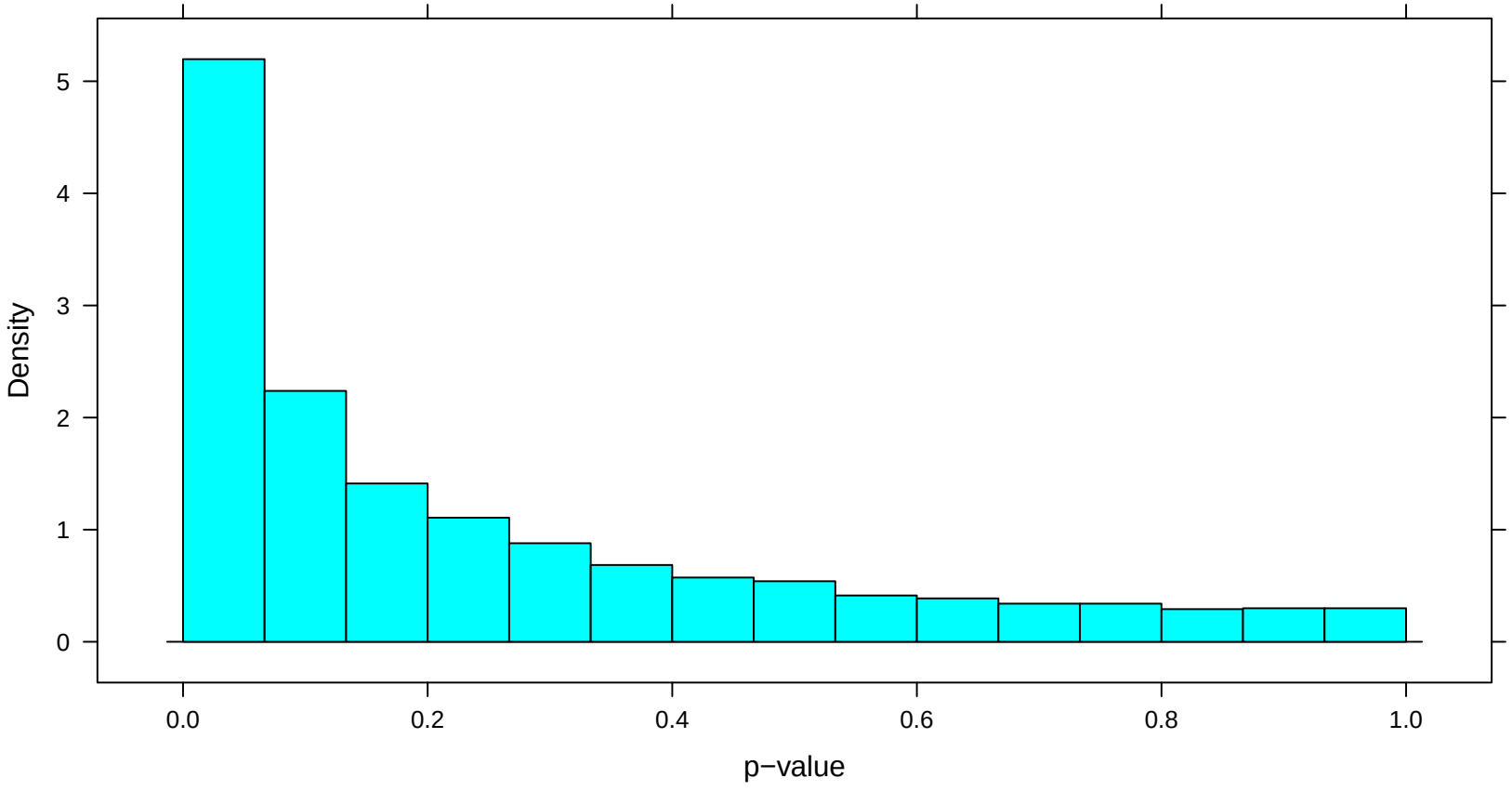
Статистика критерия:

$$t = \sqrt{n} \frac{\bar{X} - a_0}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}} \sim_{H_0} t_{n-1}$$

Убедитесь, что вы получили аналогичные картинки

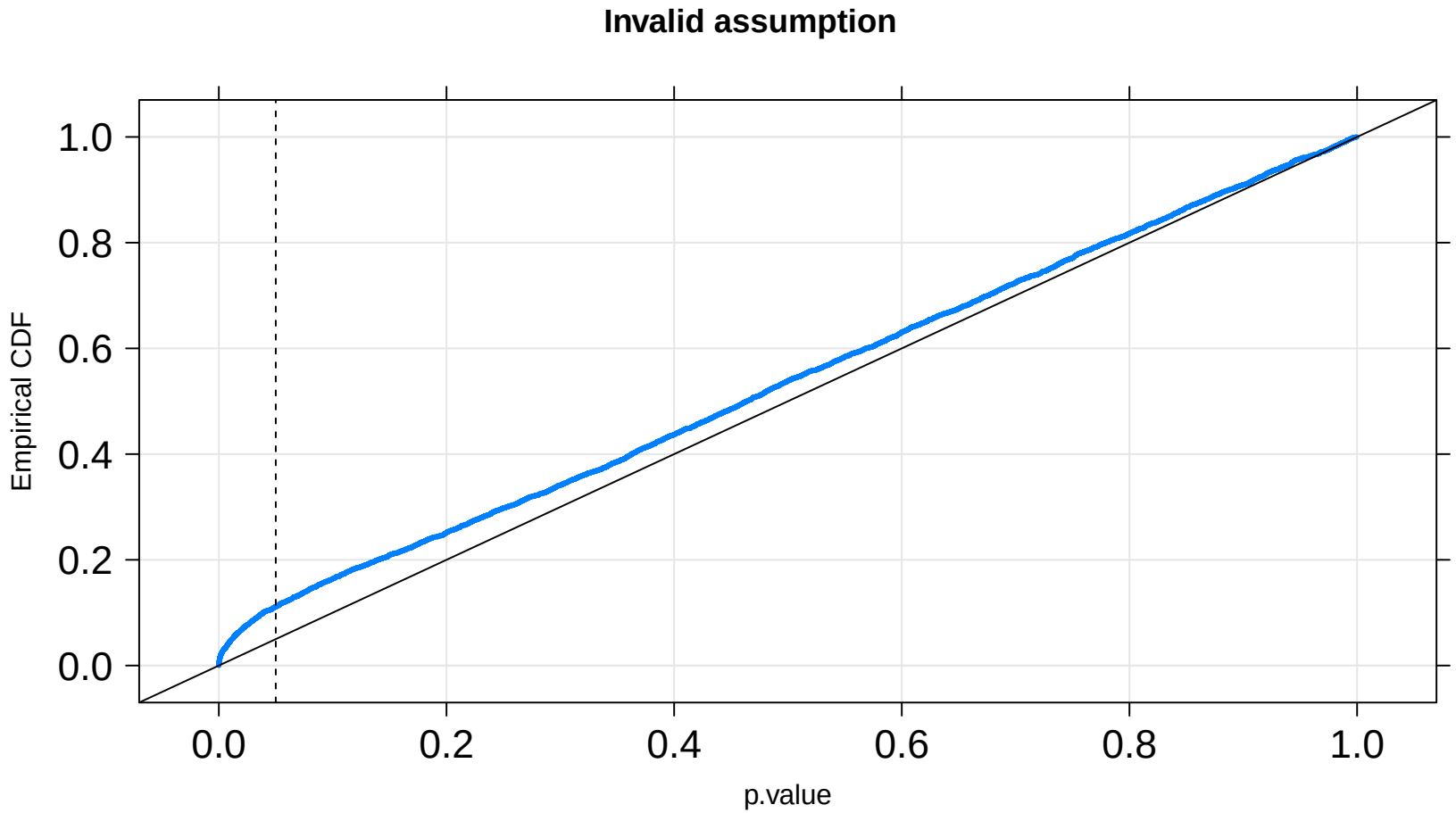
Типичные проблемы

Нарушается предположение о нормальности выборки при малом ее объеме. Пусть, например, $X_i \sim Exp(1)$ при $n = 5$.



Типичные проблемы

Нарушается предположение о нормальности выборки при малом ее объеме. Пусть, например, $X_i \sim Exp(1)$ при $n = 5$.



Одновыборочный z-критерий

Условия:

$X_1, \ldots, X_n$ — выборка из распределения с конечной дисперсией

Гипотеза:

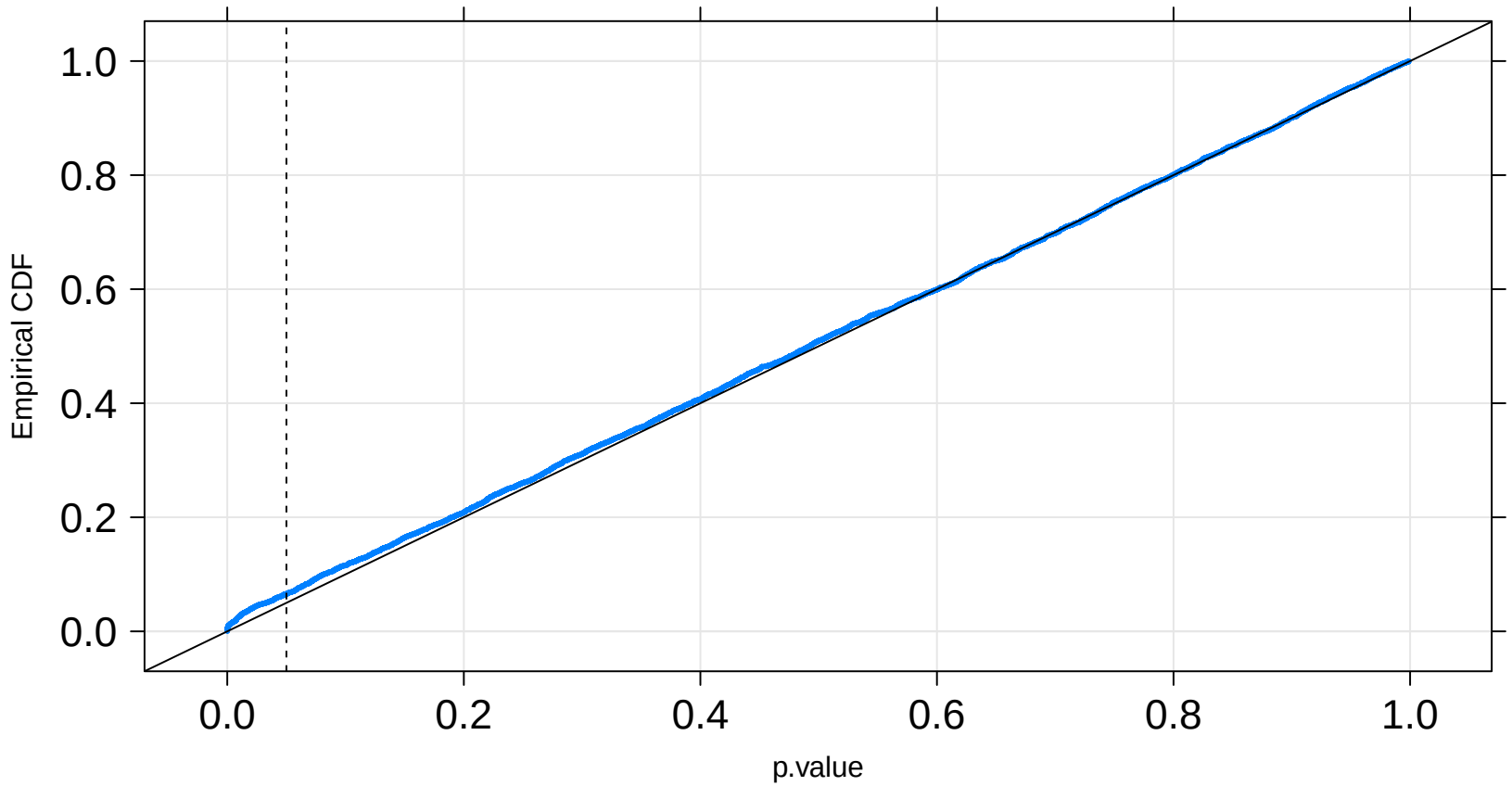
$$H_0 : EX_i = a_0 \qquad H_1 : EX_i \neq a_0$$

Статистика критерия:

$$t = \sqrt{n} \frac{\bar{X} - a_0}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}} \longrightarrow_{H_0} N(0, 1)$$

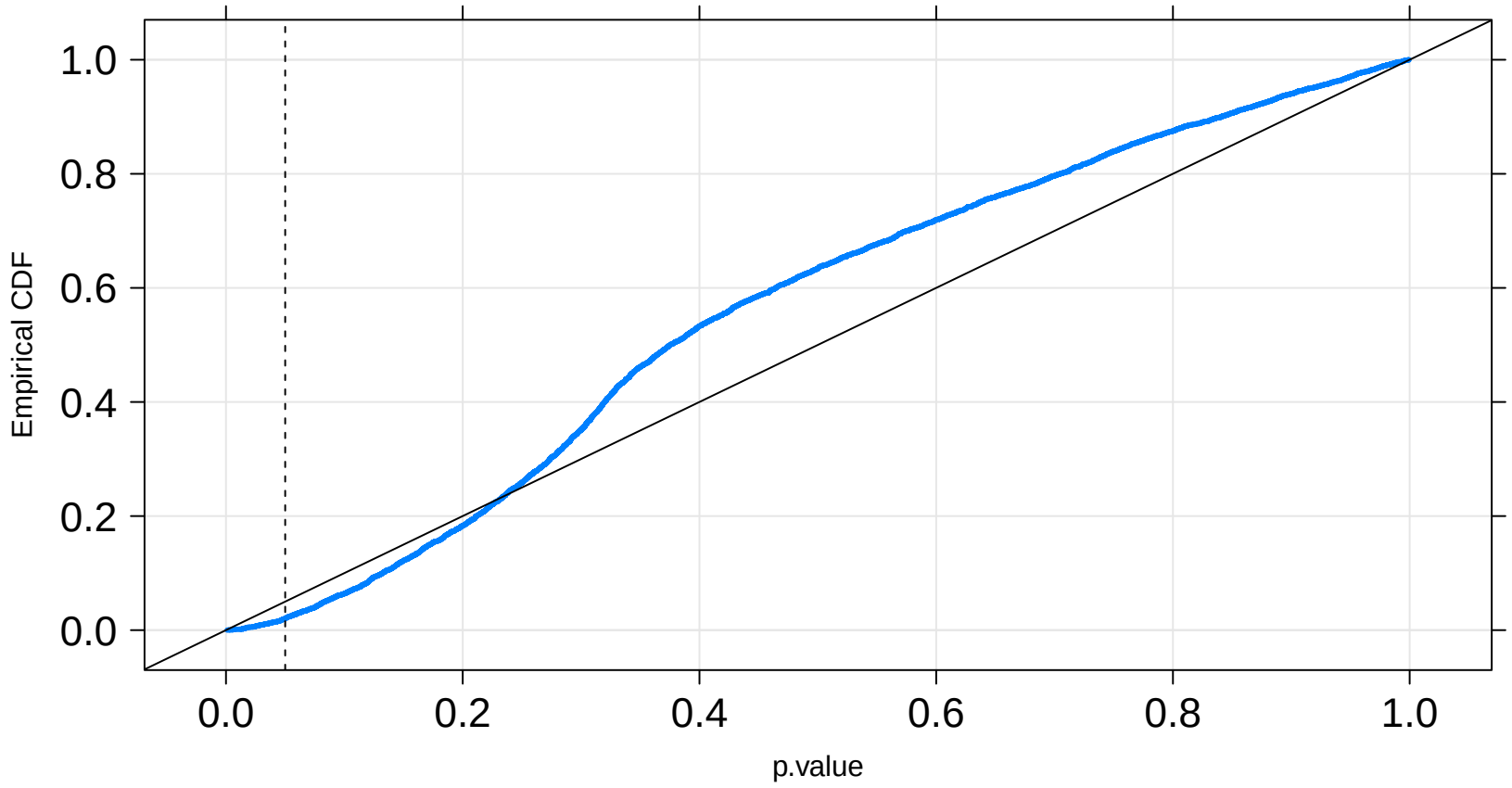
Одновыборочный z-критерий

Работает, когда распределение отлично от нормального.
$X_i \sim Exp(1)$ при $n = 50$.



Типичные проблемы

«Толстые» хвосты распределения, отсутствие дисперсии.
Пусть X_i имеют распределение Коши при $n = 1000$.



Двухвыборочный t-критерий

Условия:

$$X_1, \ldots, X_n \sim N(a_1, \theta); \qquad Y_1, \ldots, Y_m \sim N(a_2, \theta).$$

Гипотеза:

$$H_0 : a_1 = a_2 \qquad H_1 : a_1 \neq a_2$$

Статистика критерия:

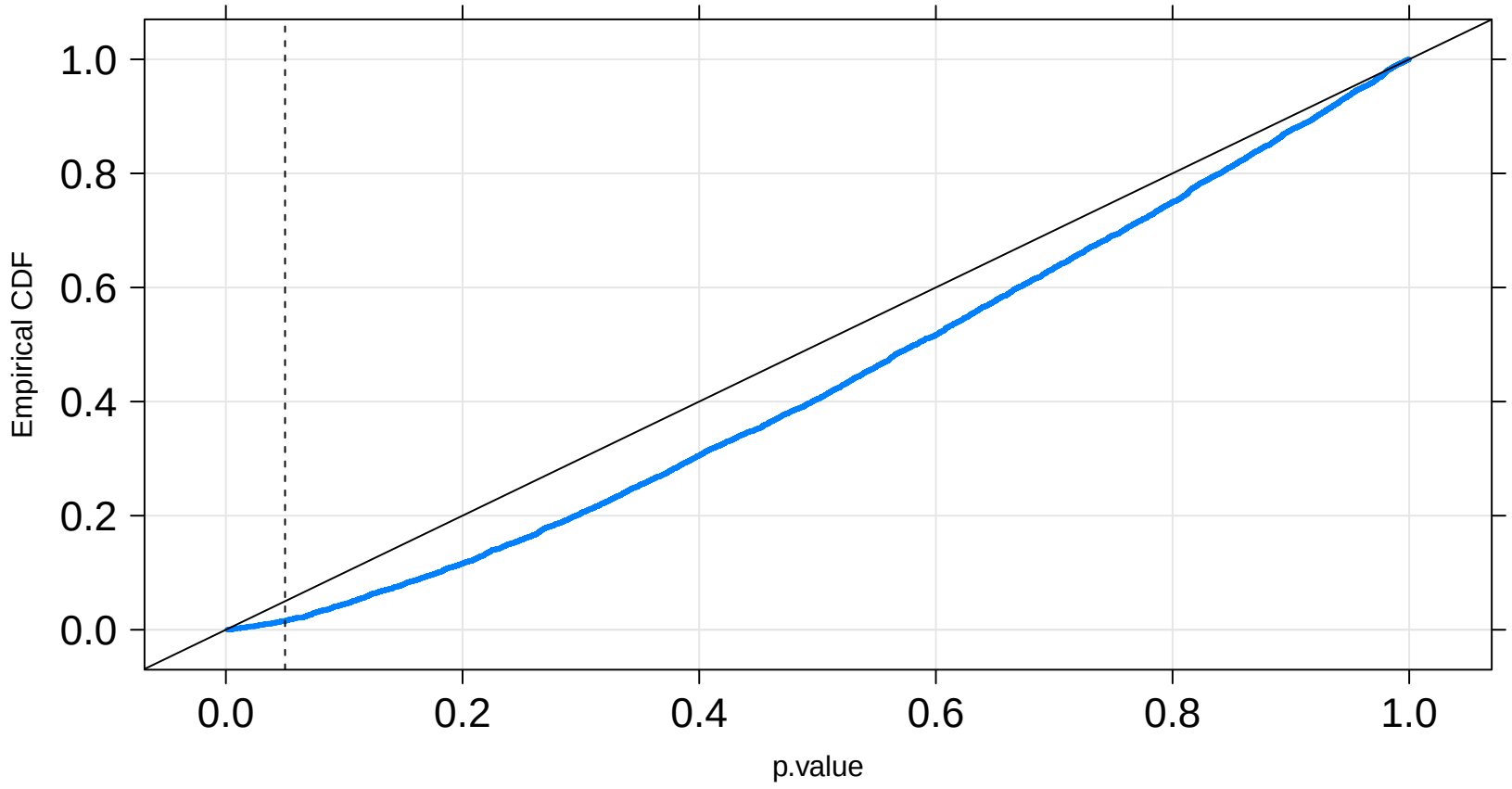
$$t = \frac{\bar{X} - \bar{Y}}{S_{X,Y}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim_{H_0} t_{n+m-2},$$

где

$$S_{X,Y} = \sqrt{\frac{(n-1)\bar{S}_X^2 + (m-1)S_Y^2}{n+m-2}}.$$

Типичные проблемы

Неодинаковая дисперсия при разных объемах выборки. Пусть $X_i \sim N(1, 1^2), Y_j \sim N(1, 2^2)$ при $n = 50, m = 100$.



Двухвыборочный t-критерий (Welch t-test)

Условия:

$$X_1, \ldots, X_n \sim N(a_1, \theta_1); \qquad Y_1, \ldots, Y_m \sim N(a_2, \theta_2).$$

Гипотеза:

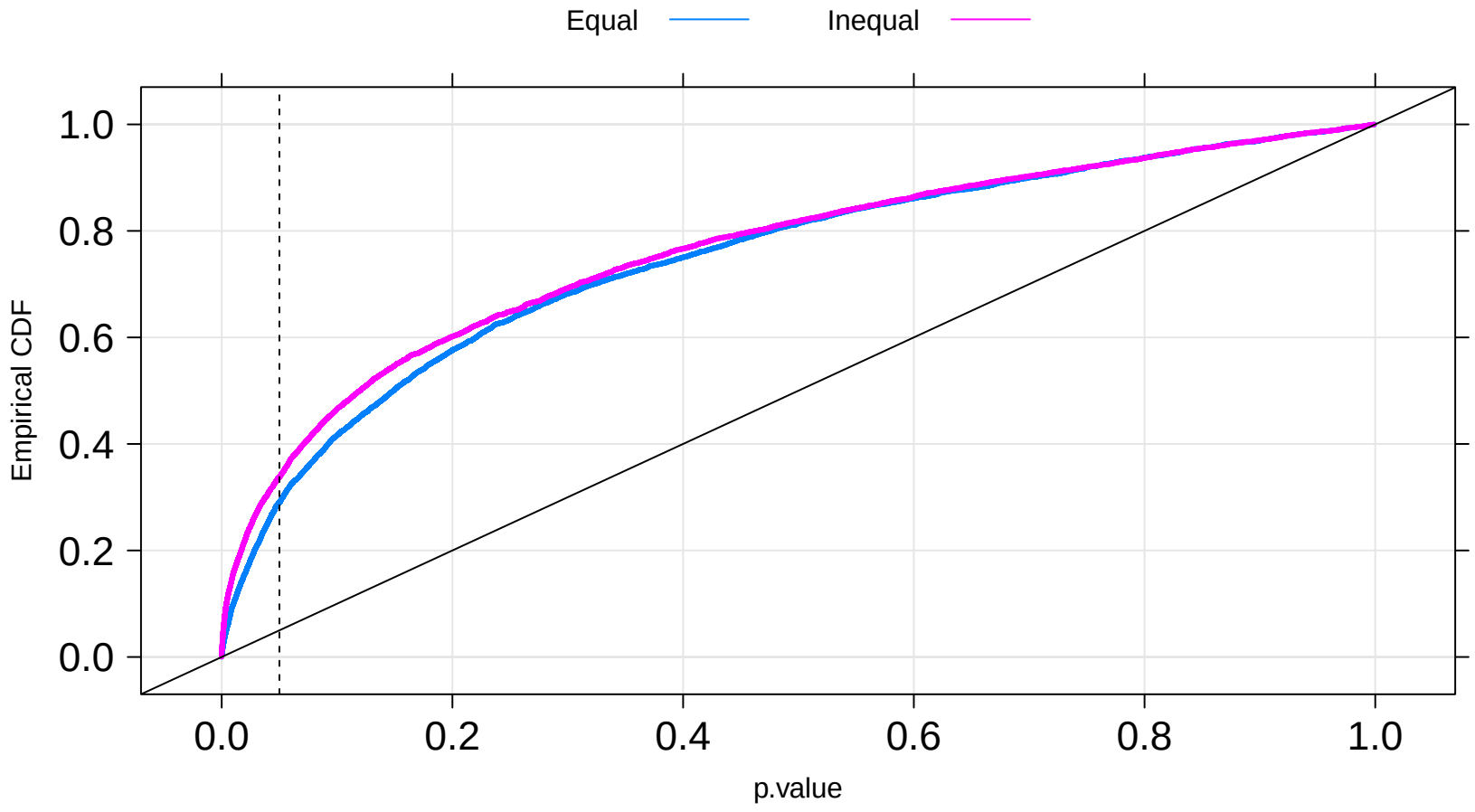
$$H_0 : a_1 = a_2 \qquad H_1 : a_1 \neq a_2$$

Статистика критерия:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\bar{S}_X^2}{n} + \frac{\bar{S}_Y^2}{m}}} \sim_{H_0} t_\nu.$$

Есть ли причина не использовать Welch t-test?

Почти никогда нет: пусть $X_i \sim N(1, 2^2), Y_j \sim N(2, 2^2)$ при $n = 10, m = 2000$.



Двухвыборочный z-критерий

Условия:

$X_1, \ldots, X_n; Y_1, \ldots, Y_m$ — независимые, с конечной дисперсией

Гипотеза:

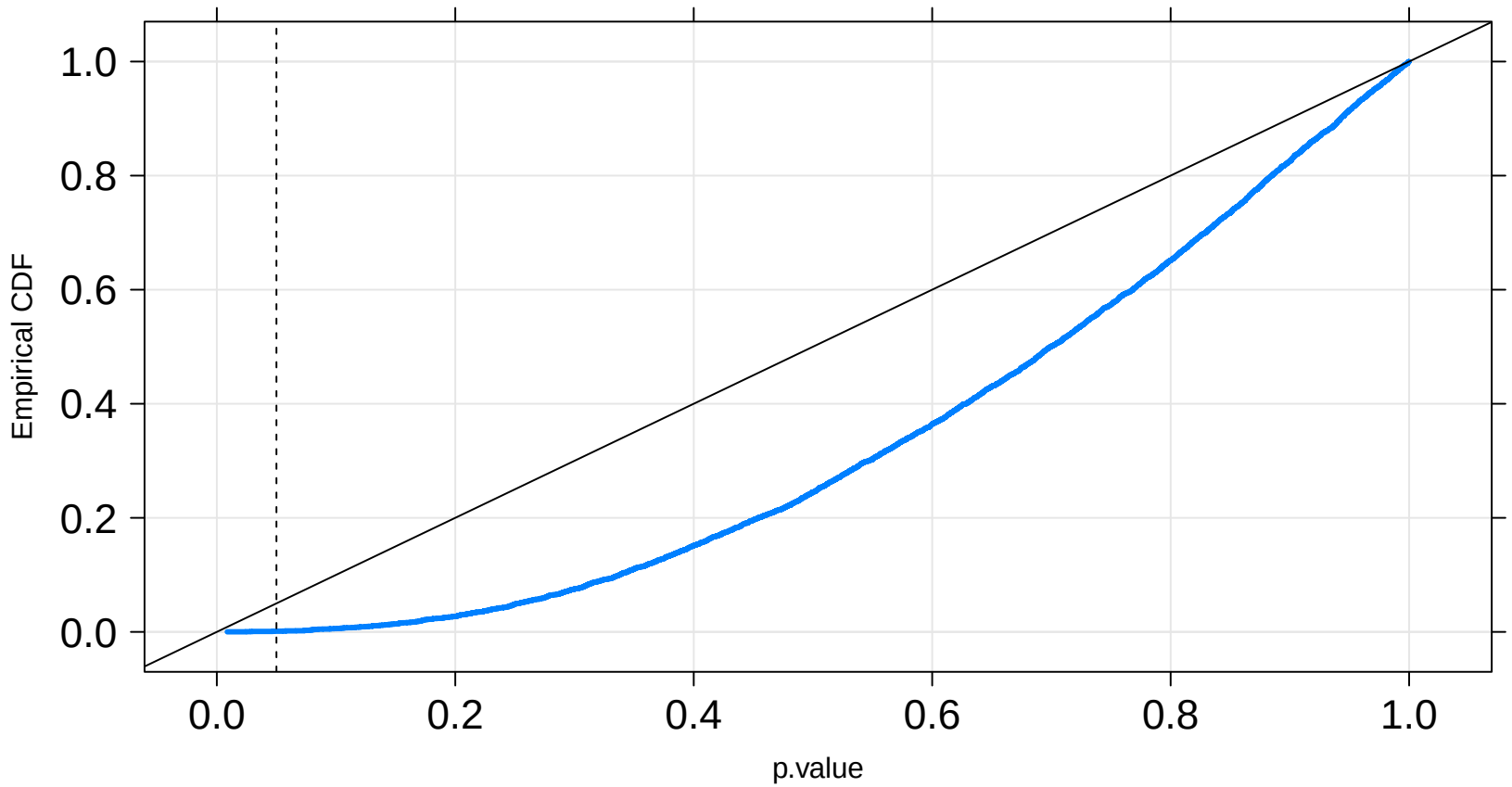
$$H_0 : EX_i = EY_j \qquad H_1 : EX_i \neq EY_j$$

Статистика критерия:

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\bar{S}_X^2}{n} + \frac{\bar{S}_Y^2}{m}}} \to_{H_0} N(0, 1).$$

Типичные проблемы

Зависимость X и Y: $X \sim N(1,1), Y - X \sim N(0,1)$



t- и z-критерии в R

Функция `t.test`:

- Одновыборочный t- и z- критерии
- Двухвыборочные t- и z- критерии
- Одинаковая / разные дисперсии: параметр `var.equal`
- Парный t-критерий
- Доверительные интервалы для среднего / разности в средних

Критерий Манна-Уитни

Условия:

$X_1, \ldots, X_n; Y_1, \ldots, Y_m$ — независимые, есть плотность

Гипотеза:

$$H_0 : \mathcal{L}(X_i) = \mathcal{L}(Y_j) \qquad H_1 : \mathcal{L}(X_i) \neq \mathcal{L}(Y_j)$$

Статистика критерия:

$$U = R_1 - \frac{n(n+1)}{2},$$

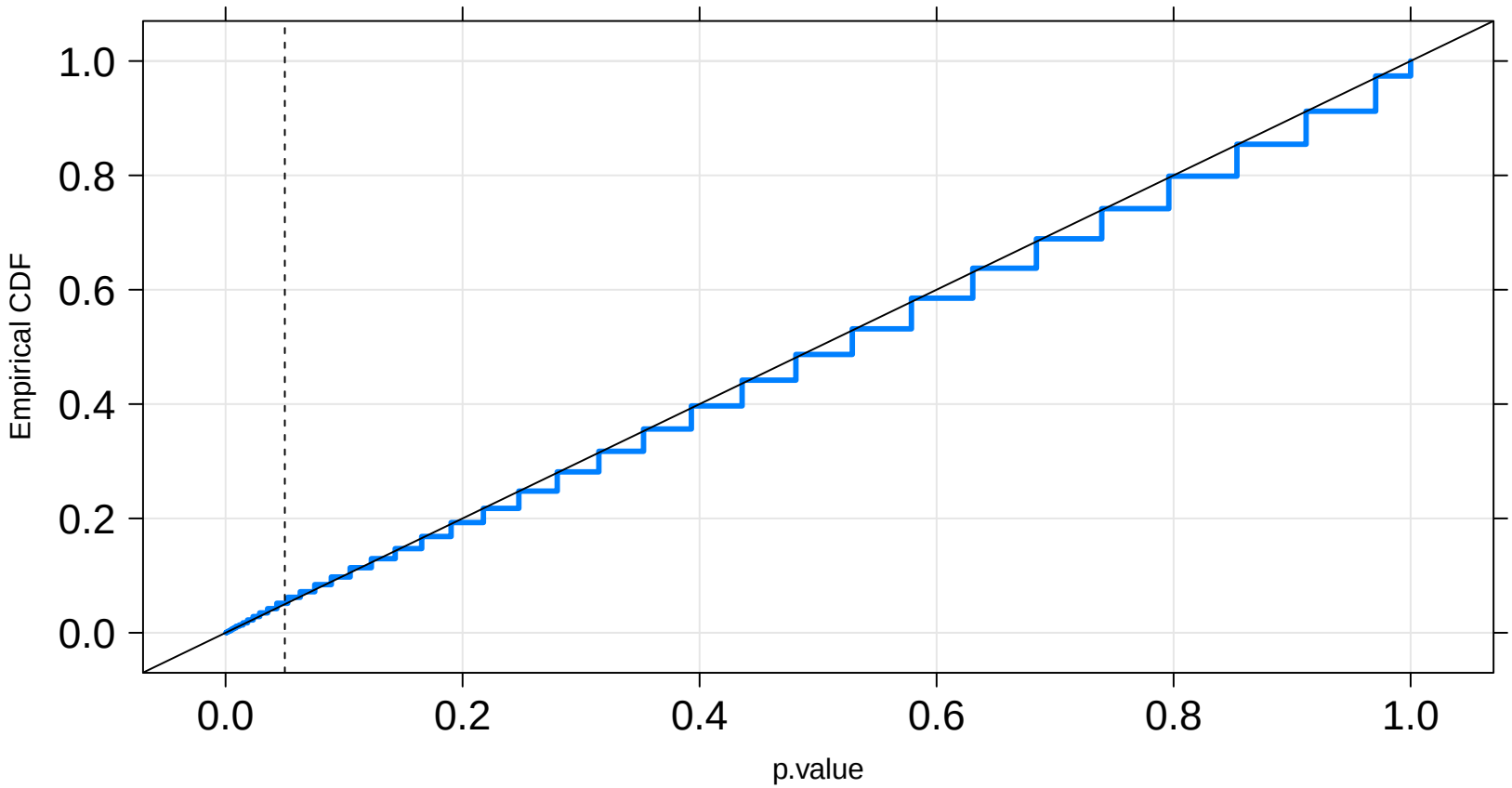
где R_1 — сумма рангов X_i в объединенной выборке.

Типичные проблемы

На самом деле нулевая гипотеза иная:

$$H_0 : P(X < Y) = P(Y < X) = 1/2$$

Пусть $X_i \sim N(0, 1)$, $Y_j \sim t_5$, $n = m = 10$.



Критерий Колмогорова-Смирнова

Предположения:

X_i — с абсолютно непрерывной функцией распределения F

Гипотеза:

$$H_0 : F(x) = G(x) \qquad H_1 : F(x) \neq G(x)$$

Статистика критерия:

$$D_n = \sqrt{n} \sup_x |F_n(x) - G(x)| \to_{H_0} K,$$

$F_n(x)$ — эмпирическая функция распределения $X_1, \ldots, X_n$.

Критерии типа ω^2

- Предположения аналогичны критерию Колмогорова-Смирнова.
- Статистика критерия ω^2 (Крамера-Смирнова-фон Мизеса):

$$\omega^2 = n \int_{-\infty}^{+\infty} \left(F_n(x) - G(x)\right)^2 \, dG(x)$$

- Статистика критерия Андерсона-Дарлинга:

$$AD = n \int_{-\infty}^{+\infty} \frac{\left(F_n(x) - G(x)\right)^2}{G(x)\left(1 - G(x)\right)} \, dG(x)$$

- Семейство критериев со статистикой:

$$n \int_{-\infty}^{+\infty} \left(F_n(x) - G(x)\right)^2 w(x) \, dG(x)$$

Типичные ошибки: оценка параметров

Как проверить сложную гипотезу?

$$H_0 : \mathcal{L}(X_i) \in \left\{ N(a, \sigma^2) \right\}$$

Типичные ошибки: оценка параметров

Как проверить сложную гипотезу?

$$H_0 : \mathcal{L}(X_i) \in \left\{ N(a, \sigma^2) \right\}$$

Есть большой соблазн оценить неизвестные параметры и подставить, например, в критерий Колмогорова-Смирнова $G(x, \hat{\theta}_n)$.

Типичные ошибки: оценка параметров

Как проверить сложную гипотезу?

$$H_0 : \mathcal{L}(X_i) \in \left\{ N(a, \sigma^2) \right\}$$

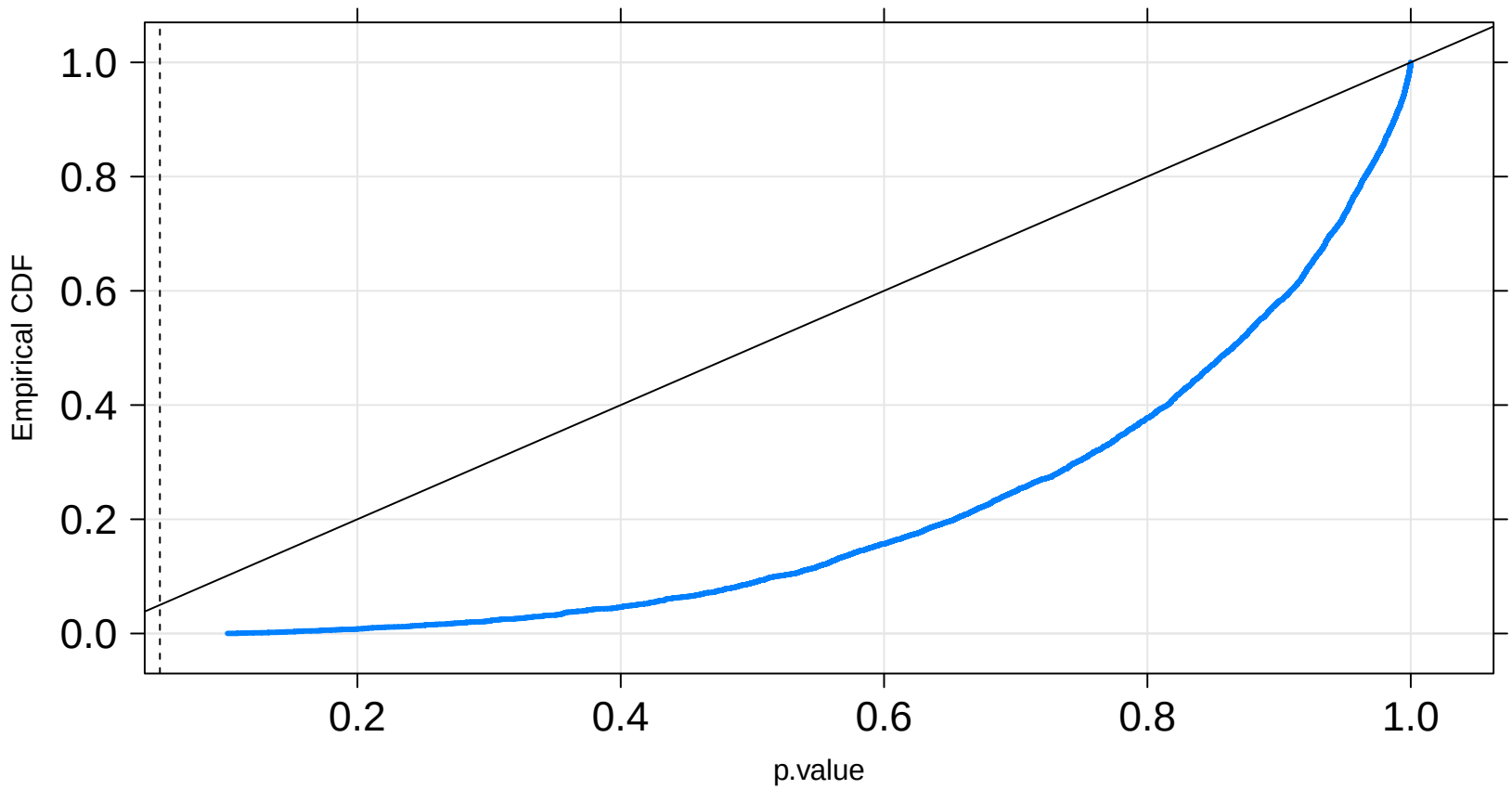
Есть большой соблазн оценить неизвестные параметры и подставить, например, в критерий Колмогорова-Смирнова $G(x, \hat{\theta}_n)$.

Не надо так!

Типичные ошибки: оценка параметров

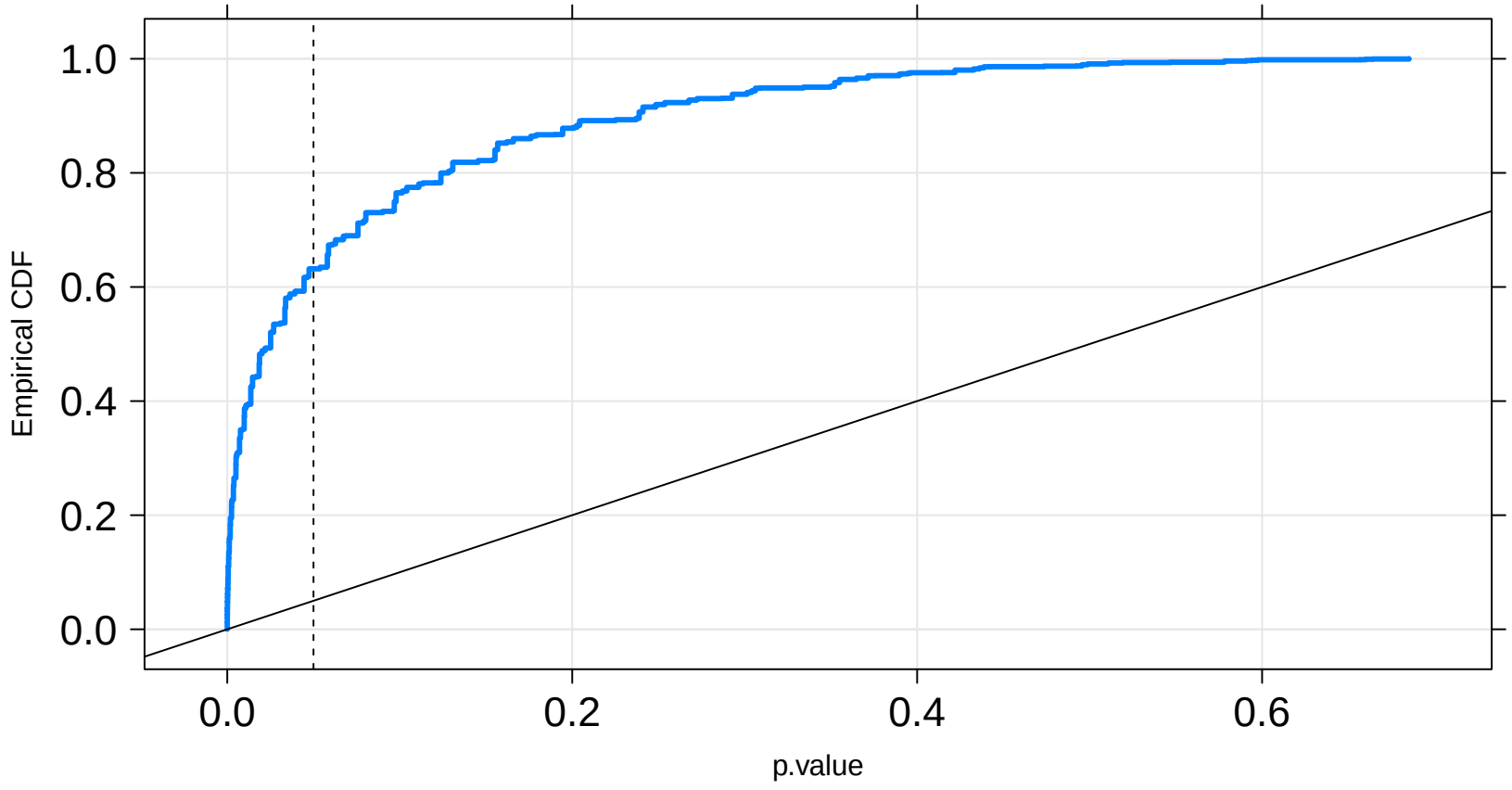
$$\sqrt{n}\sup_x \left|F_n(x) - G(x,\hat{\theta})\right| \not\to_{H_0} K,$$

$$X_i \sim N(0,1), \theta = \left(\bar{X}, \bar{S}_X^2\right):$$



Типичные ошибки: дискретные распределения

Пусть $X_i \sim \Pi(10)$



Проверка на нормальность

Специализированные критерии, учитывающие сложную гипотезу:

- Критерий Лиллиефорса
- Критерий Крамера-фон Мизеса для проверки нормальности
- ...

Пакет `nortest` в `R`.