Проверка статистических гипотез

Конкретные критерии. Типичные ошибки.

Антон Коробейников

Летняя школа по биоинформатике

28 июля 2016 года

Содержание

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:

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

В R

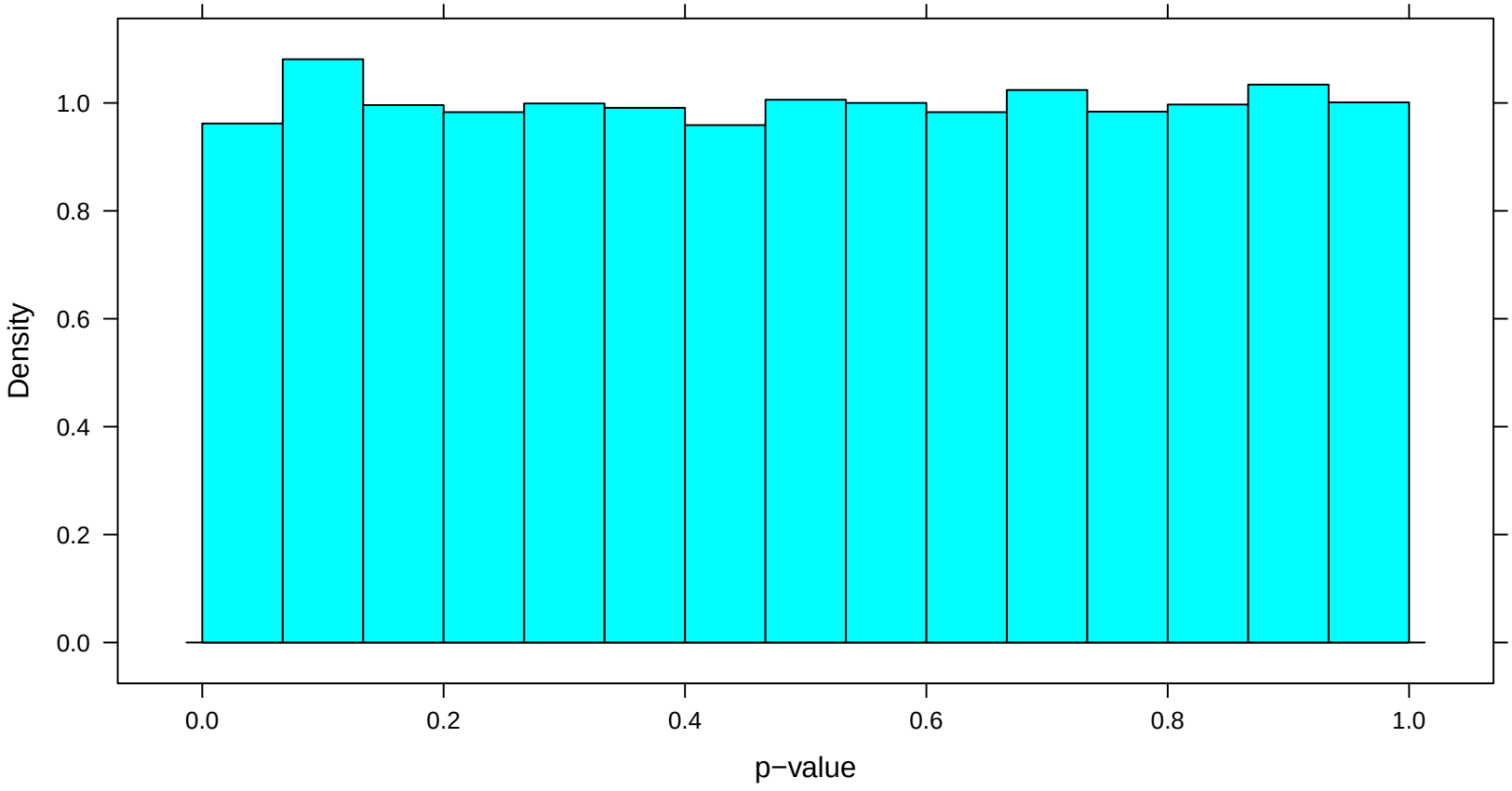
```
> v <- rnorm(10)
> t.test(v, mu = 0)

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
```

В 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
```

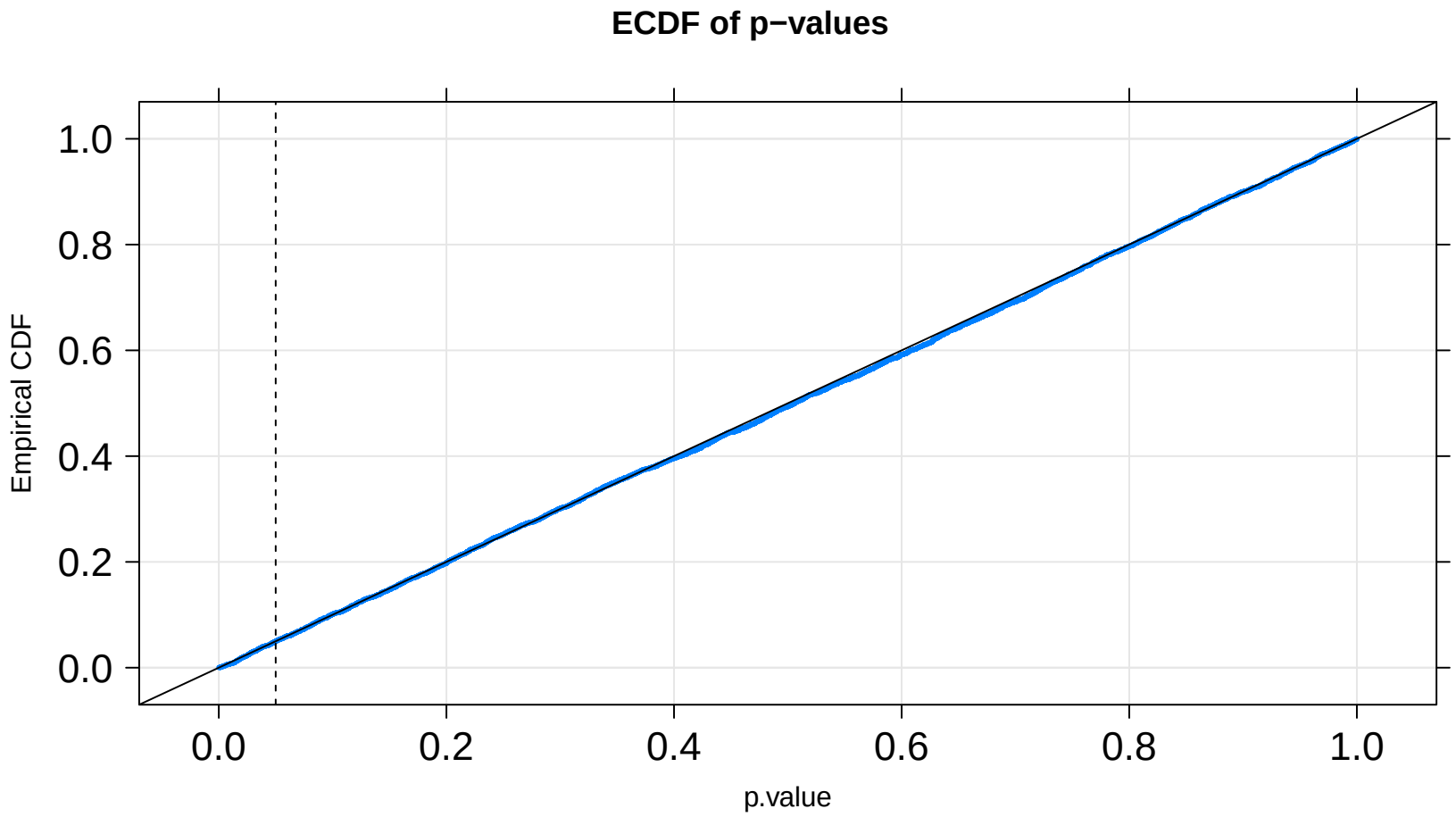
Напоминание про p-значение

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



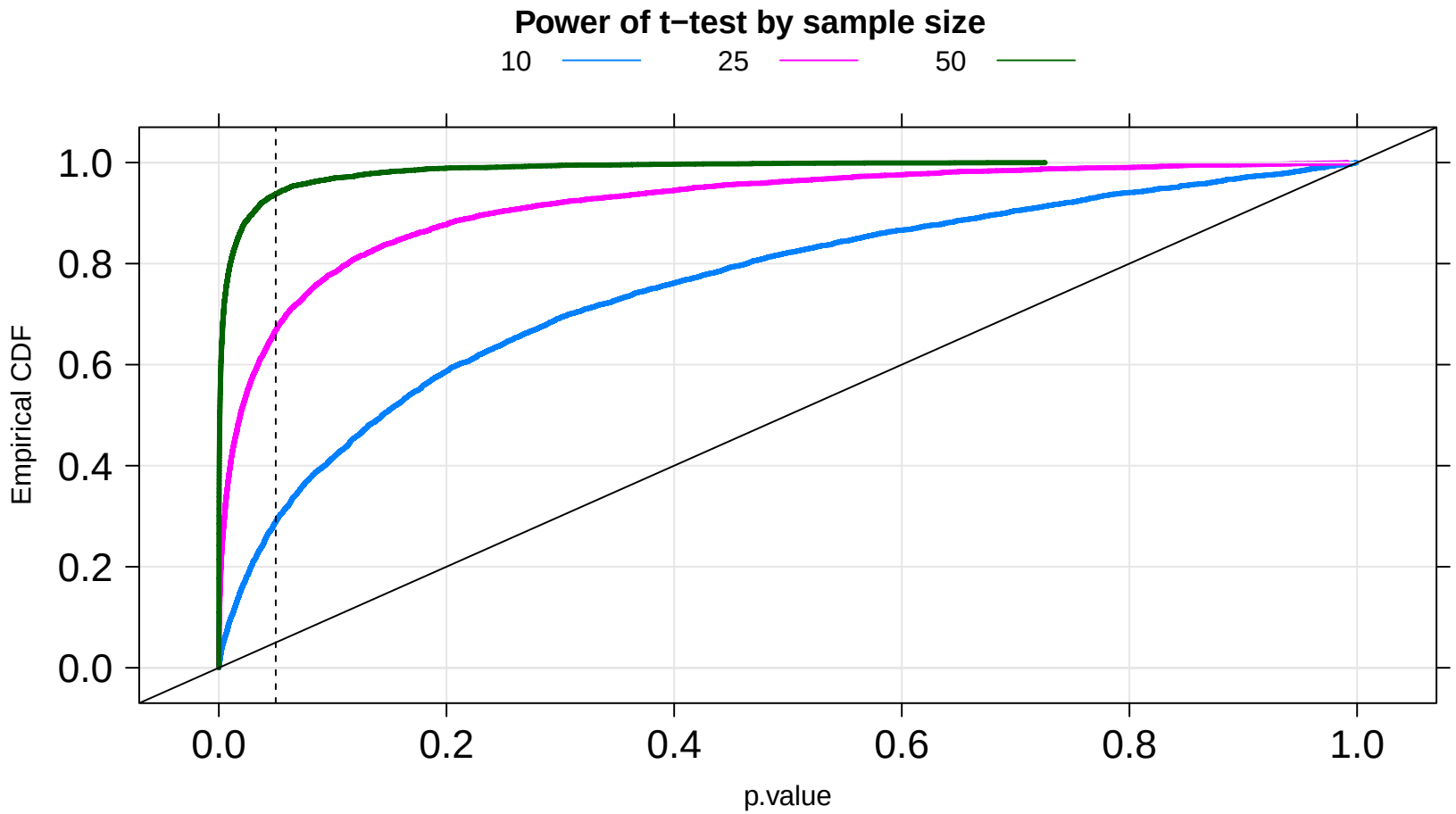
Напоминание про p-значение

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:

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

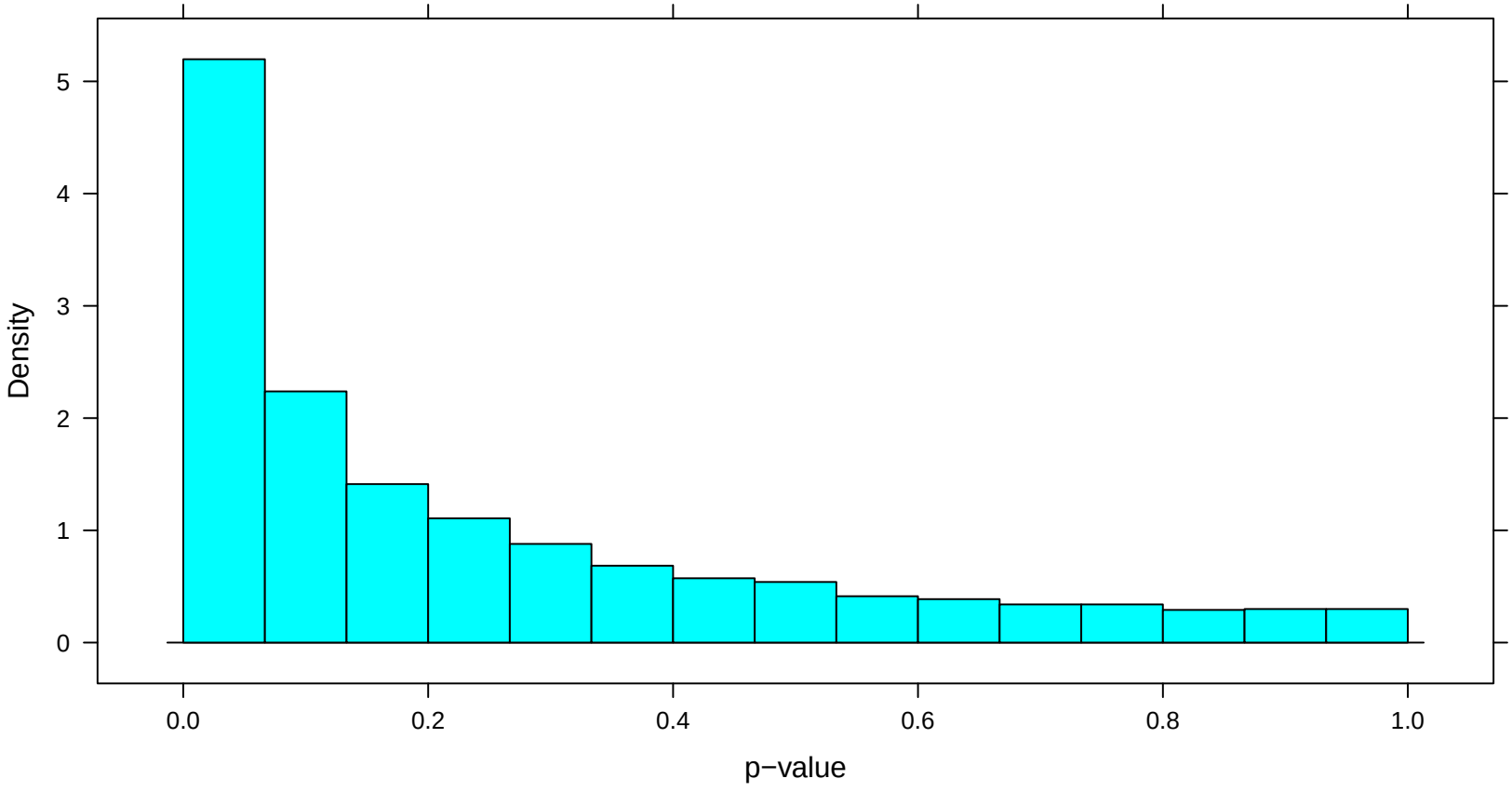
Статистика критерия:

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

Убедитесь, что вы получили аналогичные картинки

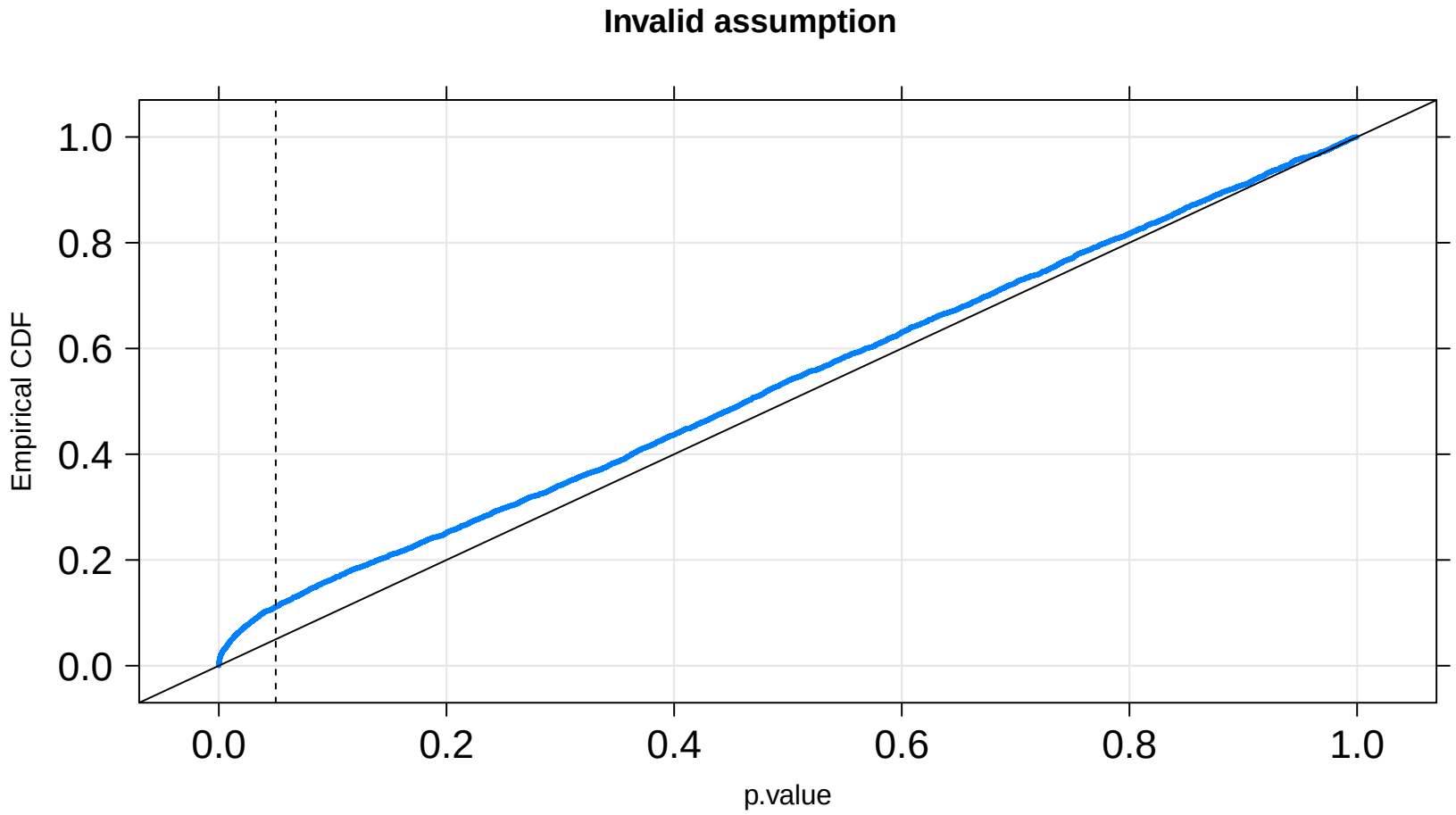
Типичные проблемы

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:

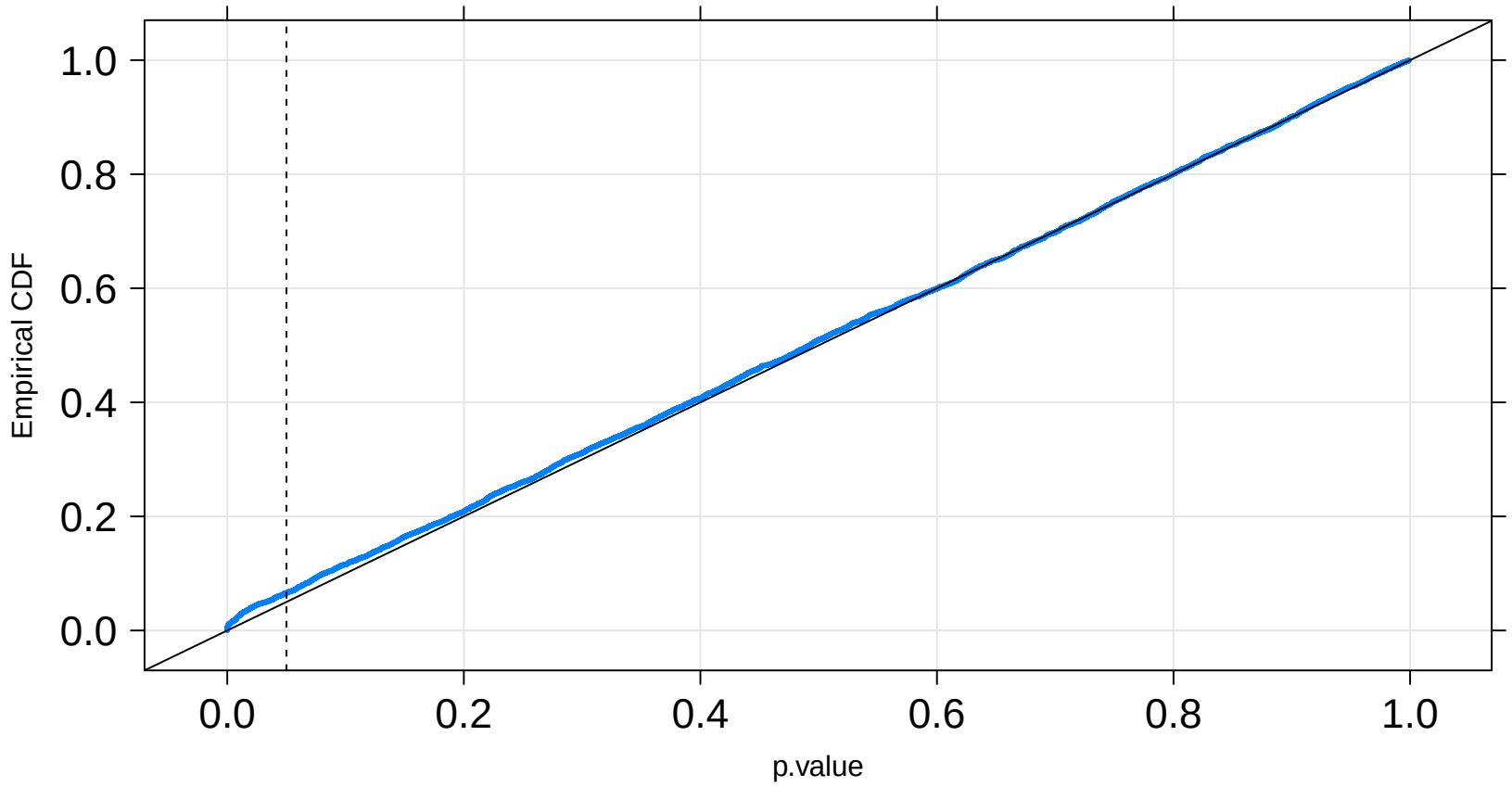
$$H_0 : EX_i = a_0 \qquad H_1 : EX_i \neq a_0$$

Статистика критерия:

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

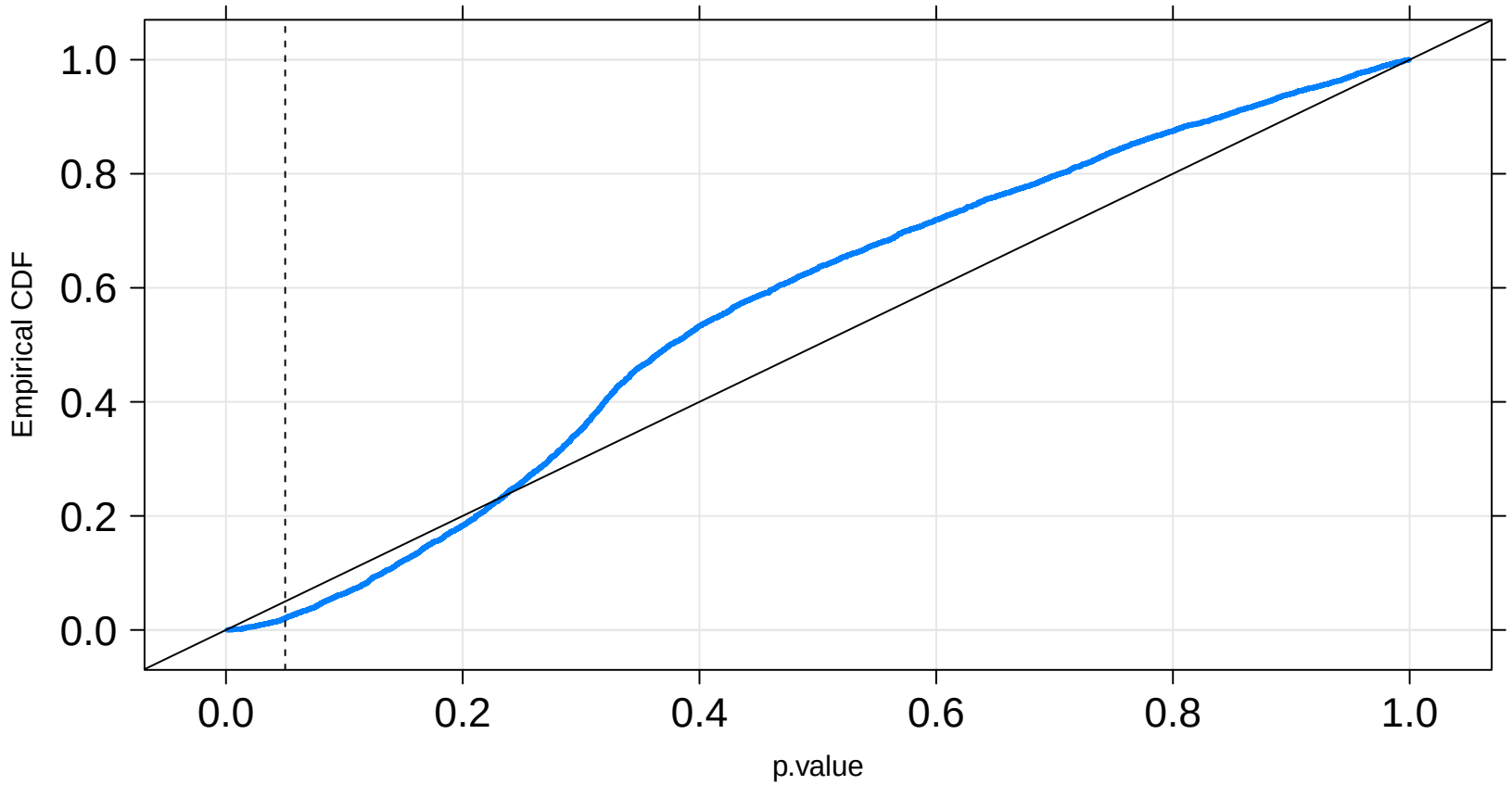
Одновыборочный z-критерий

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:

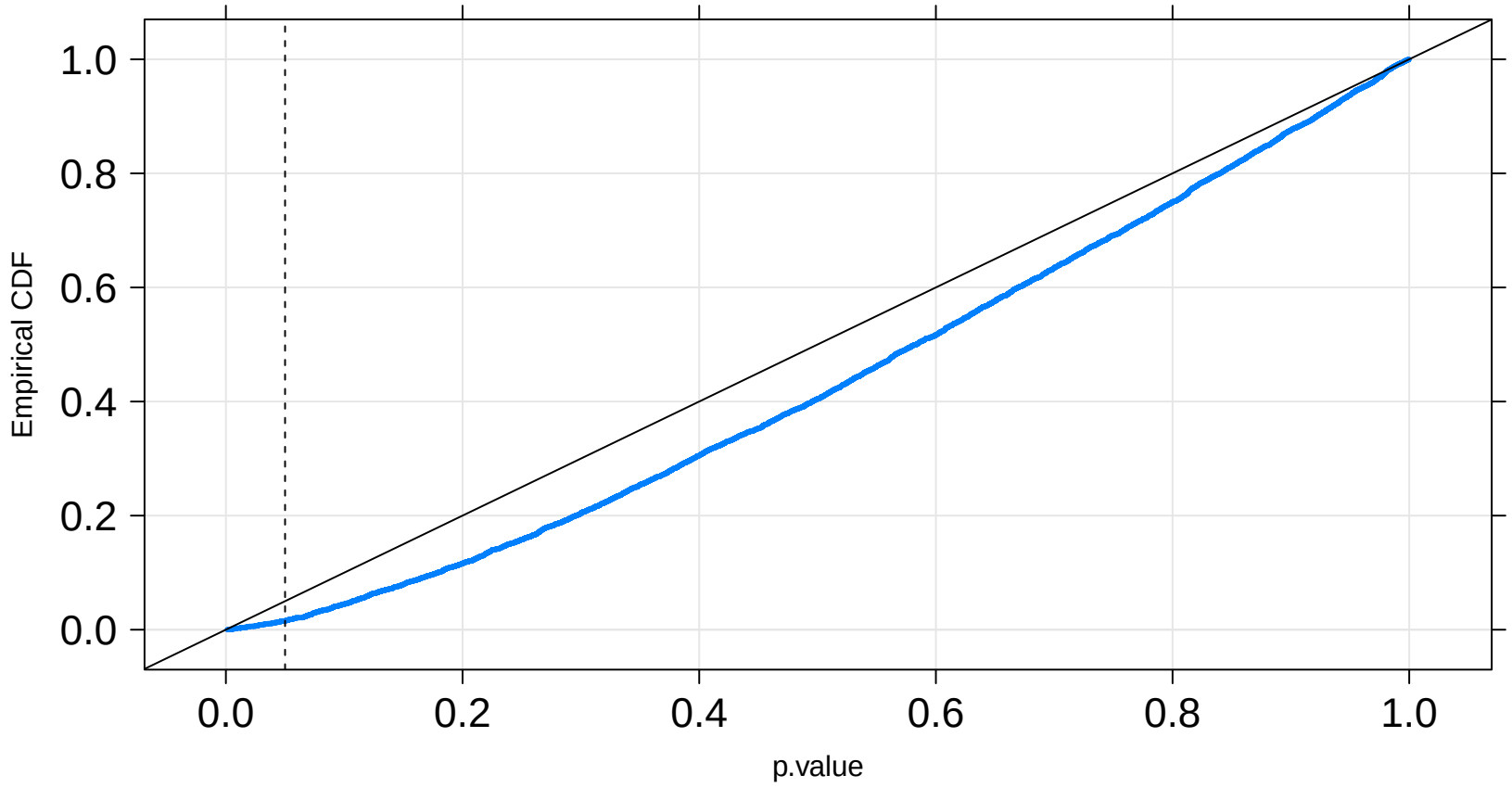
$$t = \frac{\bar{X} - \bar{Y}}{S_{X,Y}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim_{H_0} t_{n+m-2},$$

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:

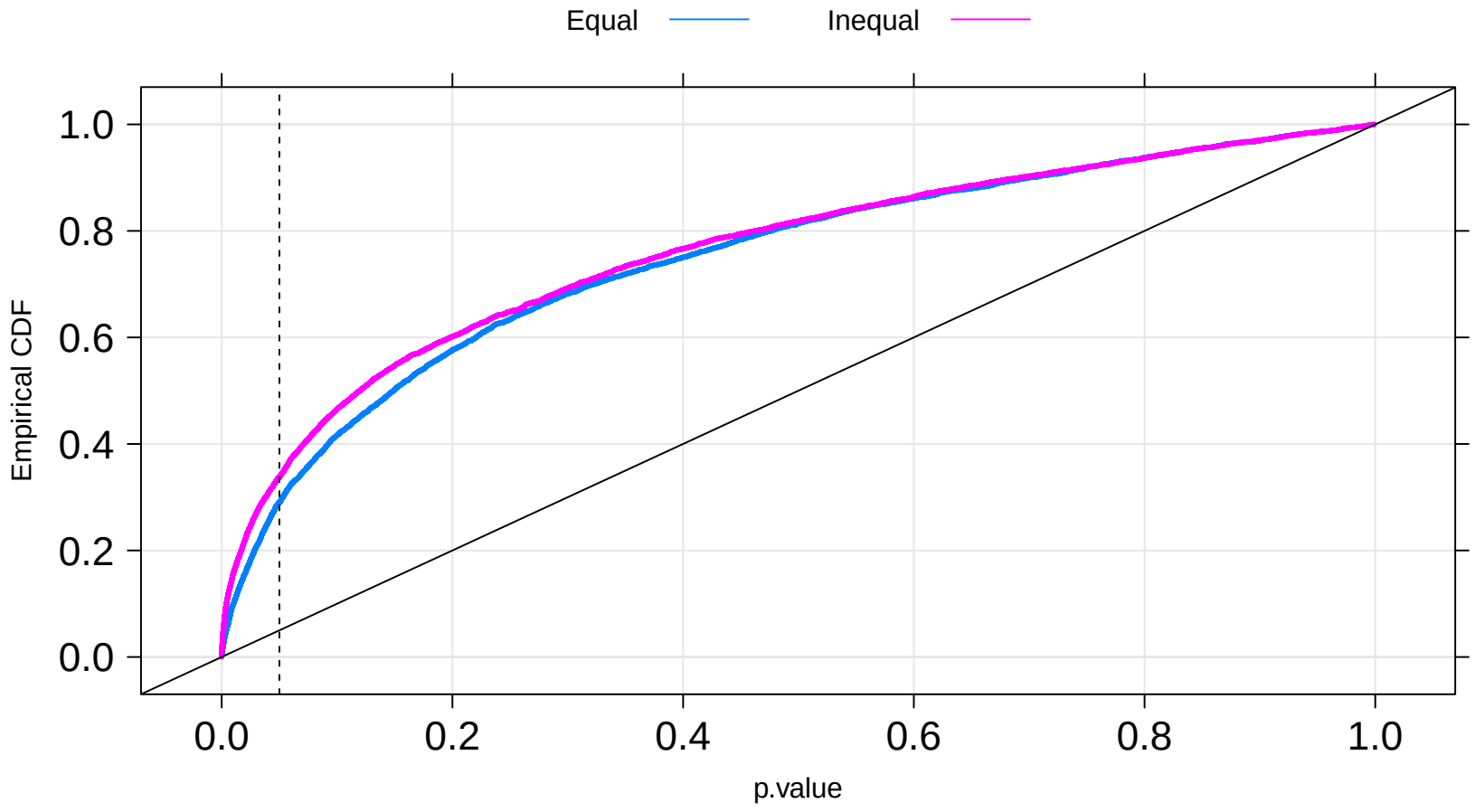
$$H_0 : a_1 = a_2 \qquad H_1 : a_1 \neq a_2$$

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:

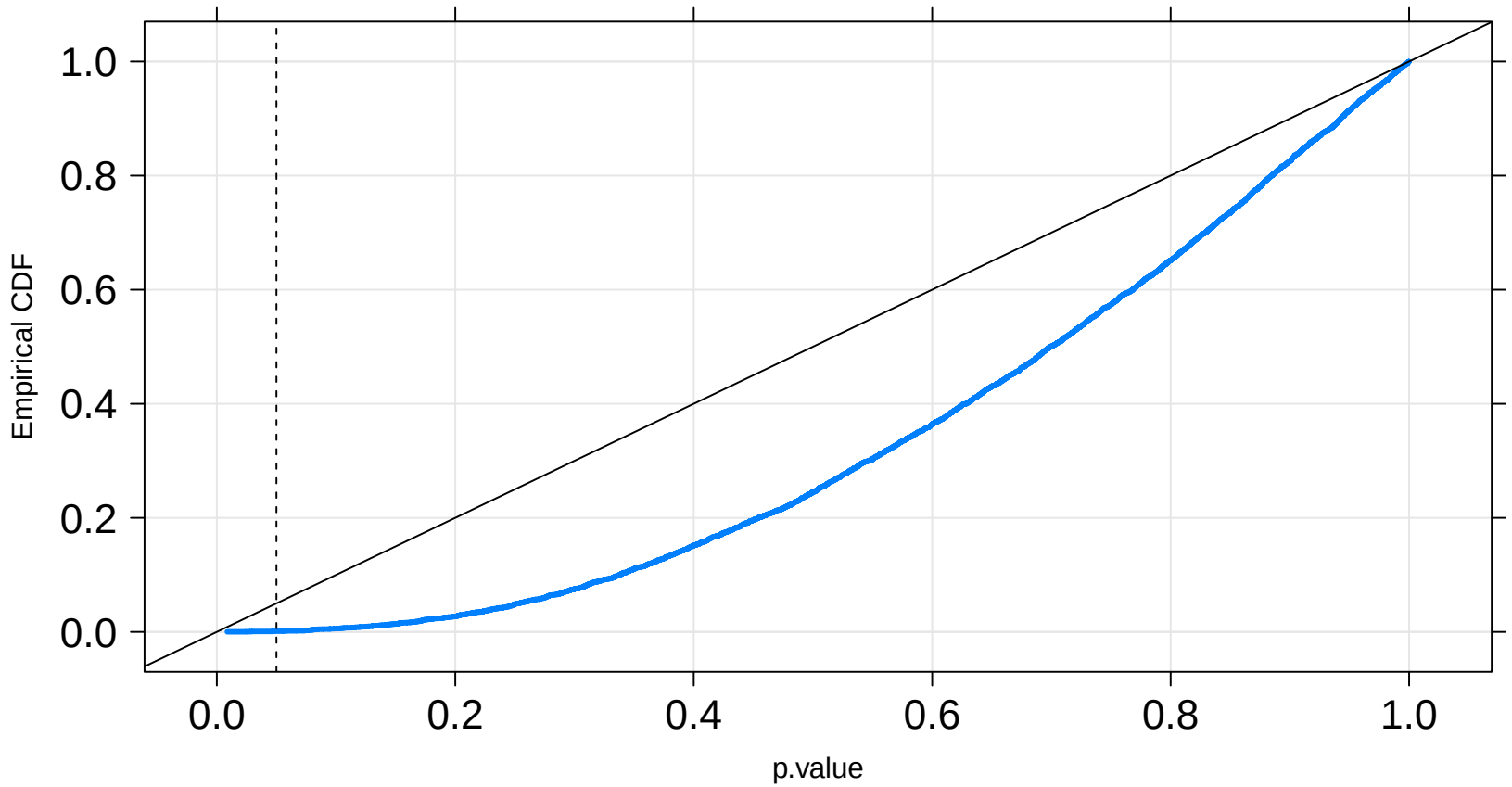
$$H_0 : EX_i = EY_j \qquad H_1 : EX_i \neq EY_j$$

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

Непараметрические критерии

- Что делать, если мы не можем ничего «разумного» предположить относительно распределения

Непараметрические критерии

- Что делать, если мы не можем ничего «разумного» предположить относительно распределения
- Качественные данные, малые объемы выборки и пр.

Непараметрические критерии

- Что делать, если мы не можем ничего «разумного» предположить относительно распределения
- Качественные данные, малые объемы выборки и пр.
- От чего-то надо отказаться...

Аналоги теста среднего

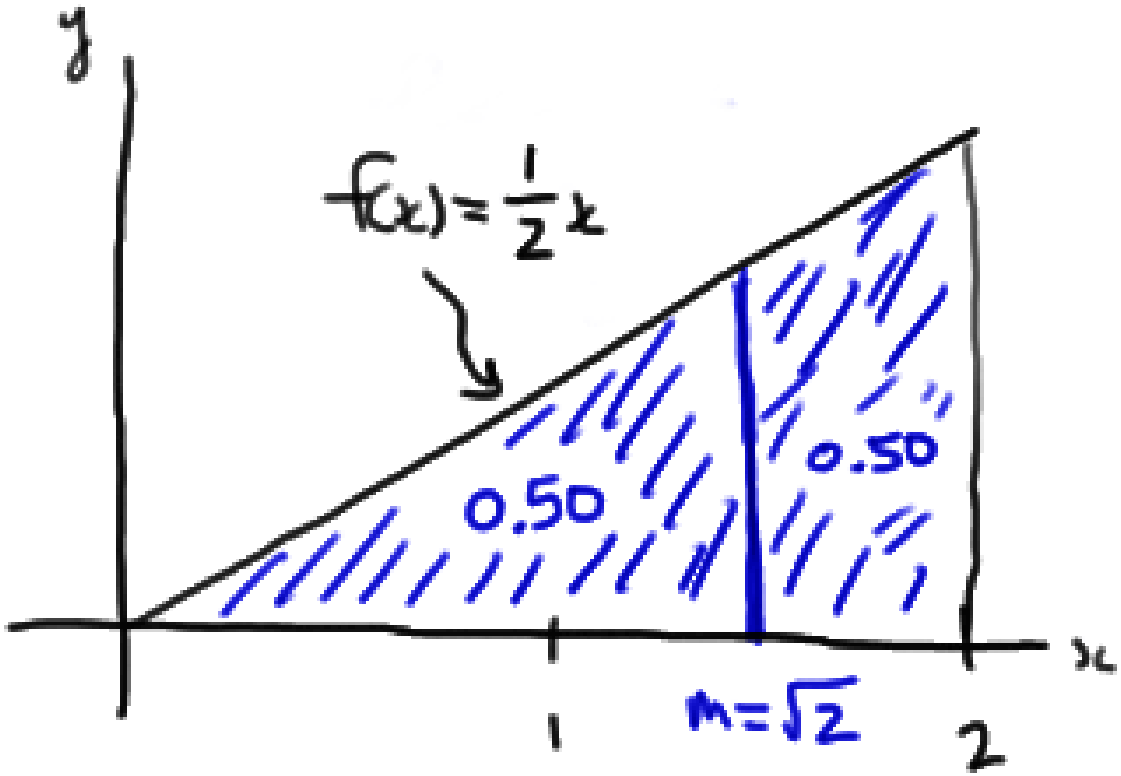
- Что нужно «уметь», чтобы считать среднее?

Аналоги теста среднего

- Что нужно «уметь», чтобы считать среднее?
- Медиана!

Аналоги теста среднего

- Что нужно «уметь», чтобы считать среднее?
- Медиана!



Критерий знаков

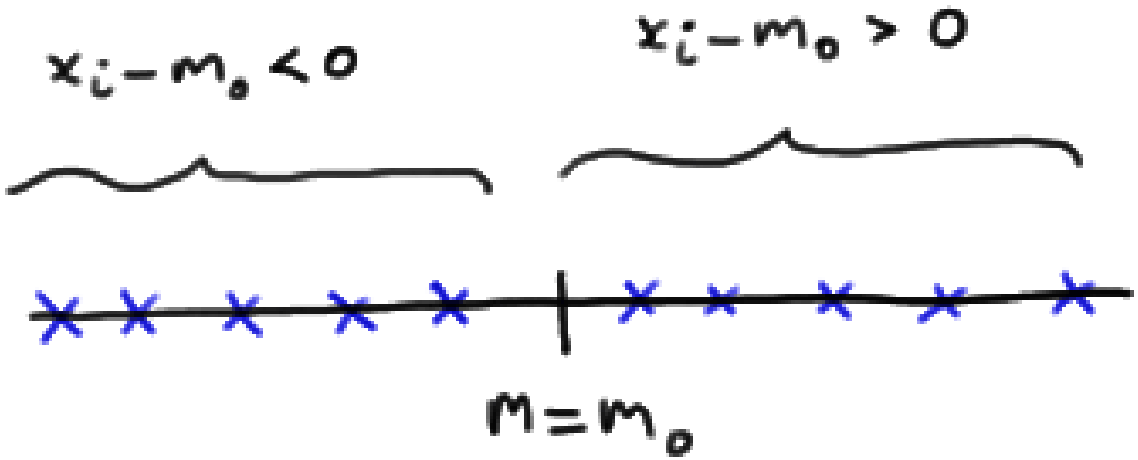
Гипотеза:

$$H_0 : \operatorname{med} X_i = m_0 \qquad H_1 : \operatorname{med} X_i \neq m_0$$

Как бы Вы строили этот критерий?

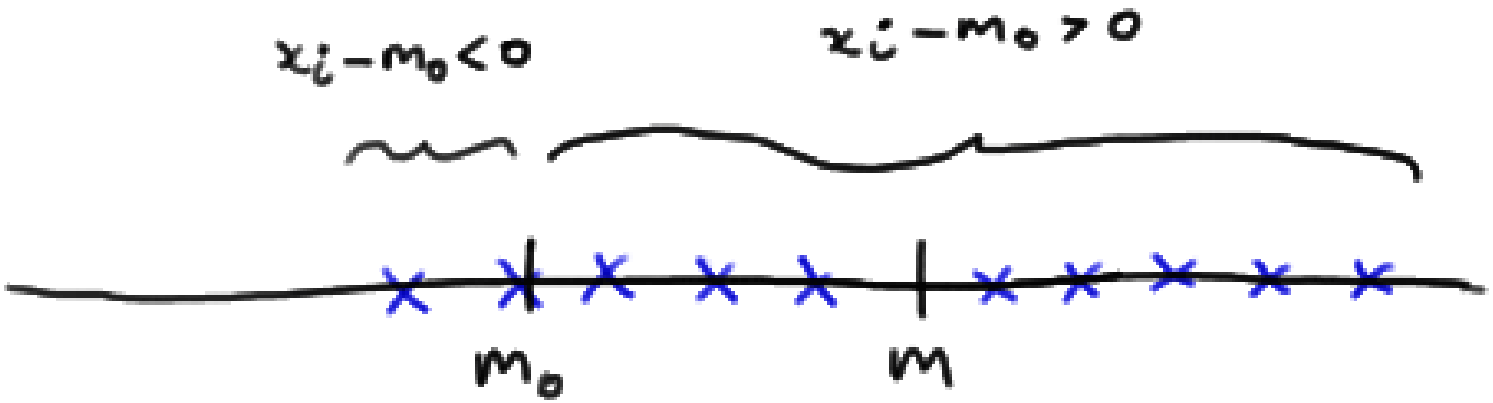
Критерий знаков

Пусть верна H_0:



Критерий знаков

Ситуация для H_1:



Критерий знаков

- Считаем $Y_i = X_i - m_0$
- Считаем N_-, N_+ — кол-во положительных и отрицательных Y_i.
- Если верна H_0, то N_i, $N_+ \sim Bin(n, \frac{1}{2})$

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