

ПРОВЕРКА ГИПОТЕЗ: НАЧАЛО

ПРЕДСКАЗАНИЕ будущего



» Как проверить?



ПРЕДСКАЗАНИЕ БУДУЩЕГО

- » Эксперимент: записываются предсказания, генерируются события, проверяется правильность предсказаний
- » $X^n = (X_1, \dots, X_n)$ — выборка результатов, например:
 - ▶ $X = 1$, если предсказание сбылось, 0 , если нет
 - ▶ X — точность предсказания (разность между фактом и прогнозом)

ПРЕДСКАЗАНИЕ БУДУЩЕГО

- » $X^n = (X_1, \dots, X_n)$ — выборка результатов,
например:
 - ▶ $X = 1$, если предсказание сбылось, 0 , если нет
 - ▶ X — точность предсказания (разность между
фактом и прогнозом)
- » Предсказатель полезен, если он предсказывает
лучше, чем генератор случайных чисел

ПРЕДСКАЗАНИЕ будущего

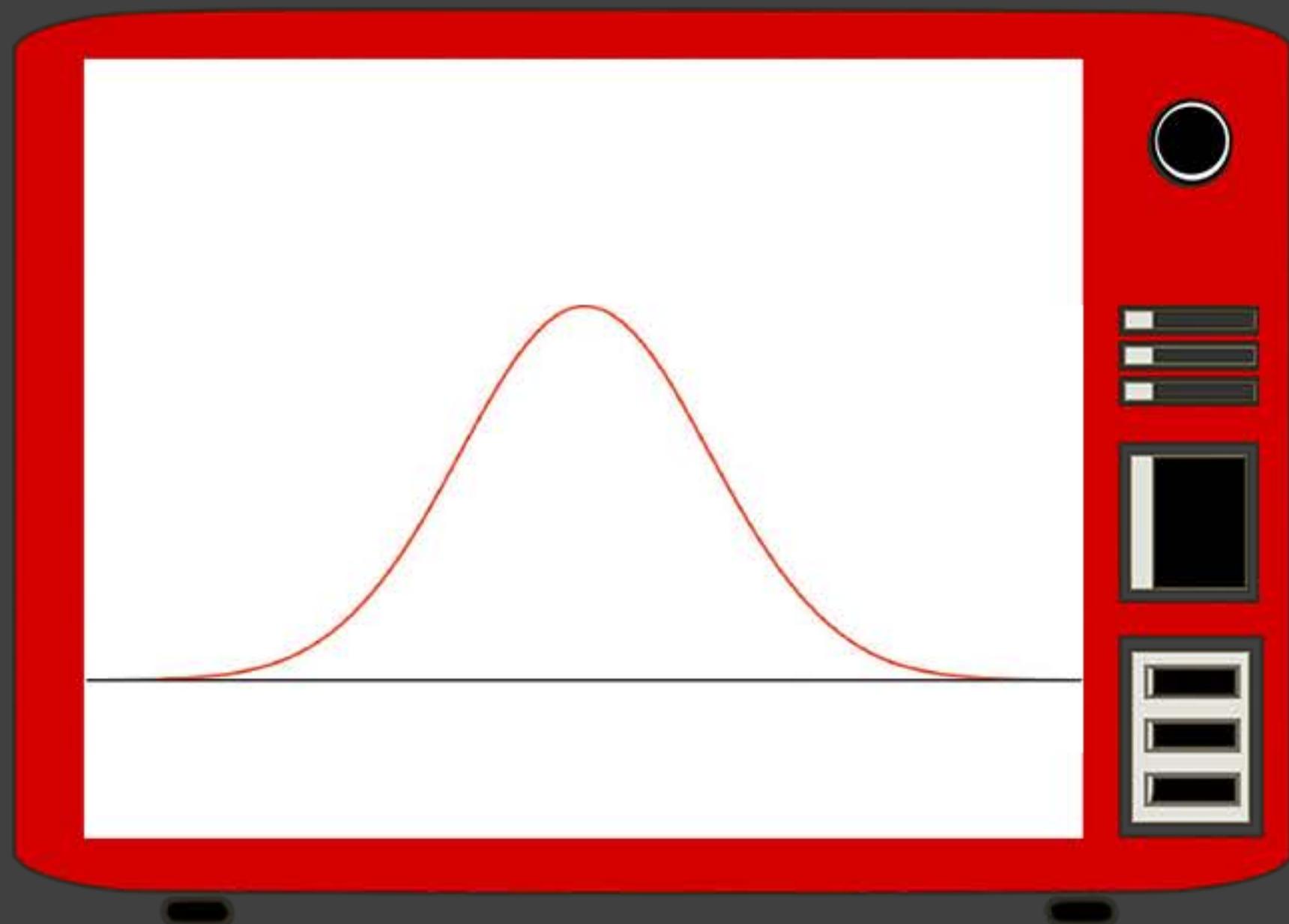
- › Предсказатель полезен, если он предсказывает лучше, чем генератор случайных чисел
- › Гипотеза: предсказатель — и есть генератор случайных чисел
- › Что говорят данные? Свидетельствуют ли они против такого предположения?

ПРОВЕРКА ГИПОТЕЗ

- » Выборка: $X^n = (X_1, \dots, X_n)$, $X \sim P$
- » Нулевая гипотеза: $H_0: P \in \omega$
- » Альтернатива: $H_1: P \notin \omega$
- » Статистика: $T(X^n)$, $T(X^n) \sim F(x)$ при H_0
 $T(X^n) \not\sim F(x)$ при H_1

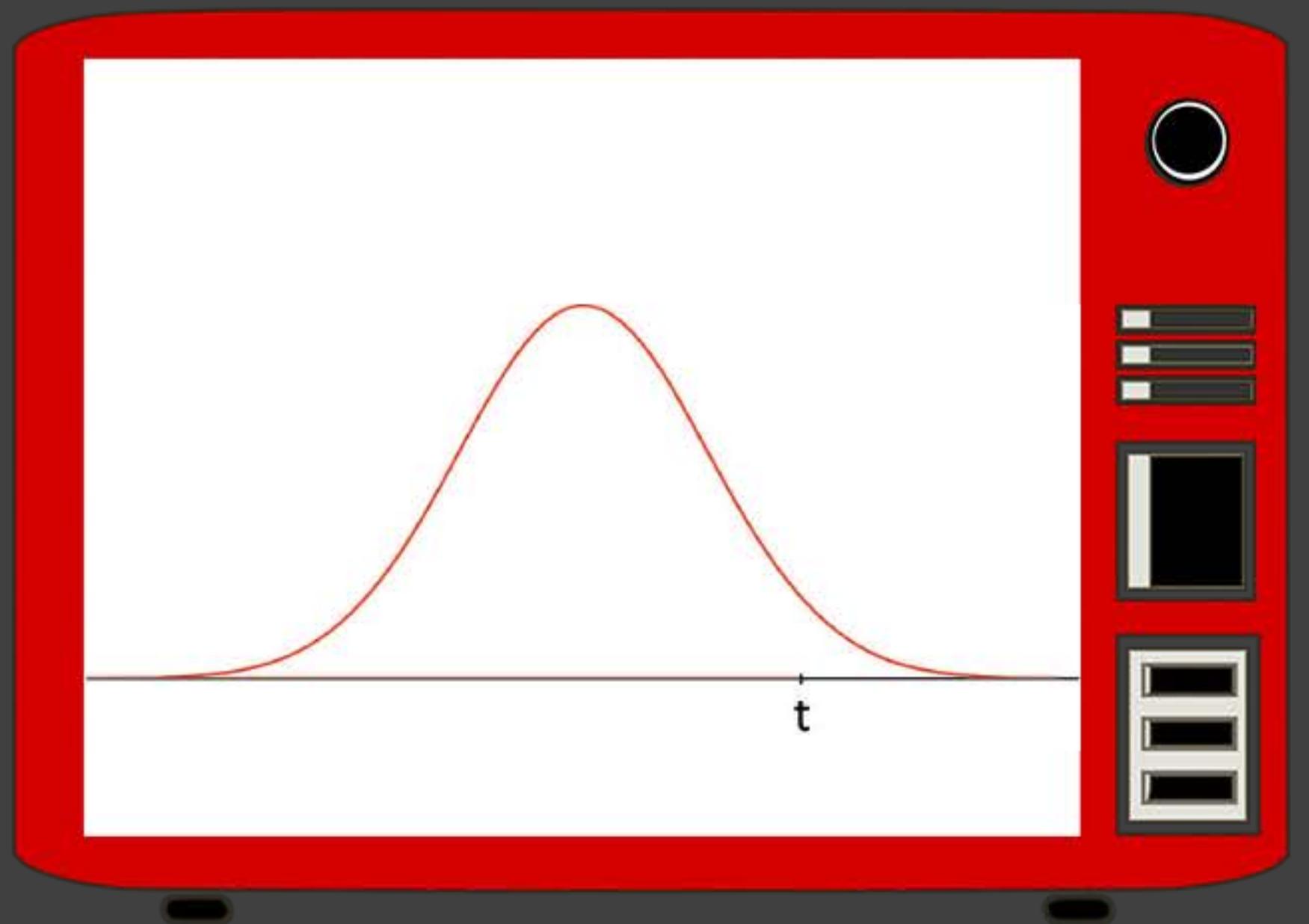
ПРОВЕРКА ГИПОТЕЗ

- » Статистика: $T(X^n)$, $T(X^n) \sim F(x)$ при H_0
 $T(X^n) \not\sim F(x)$ при H_1
- » $F(x)$ — нулевое распределение статистики
- » Вместе T и $F(x)$ — статистический критерий для проверки H_0 против H_1



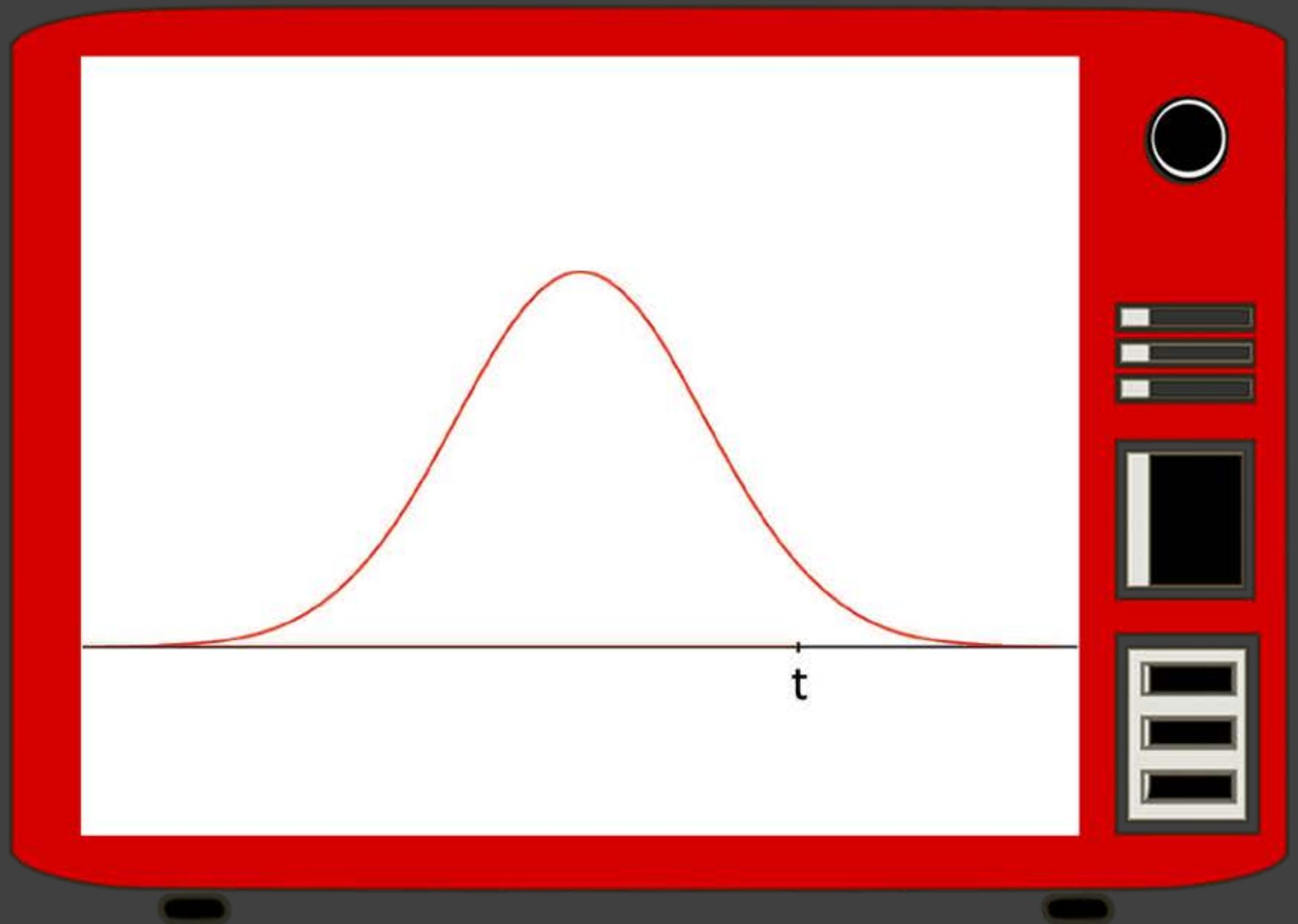
НУЛЕВОЕ РАСПРЕДЕЛЕНИЕ

- › $F(x)$ — нулевое распределение статистики
- › t — значение статистики на полученных данных.
- › Насколько оно вероятно при справедливости H_0 ?



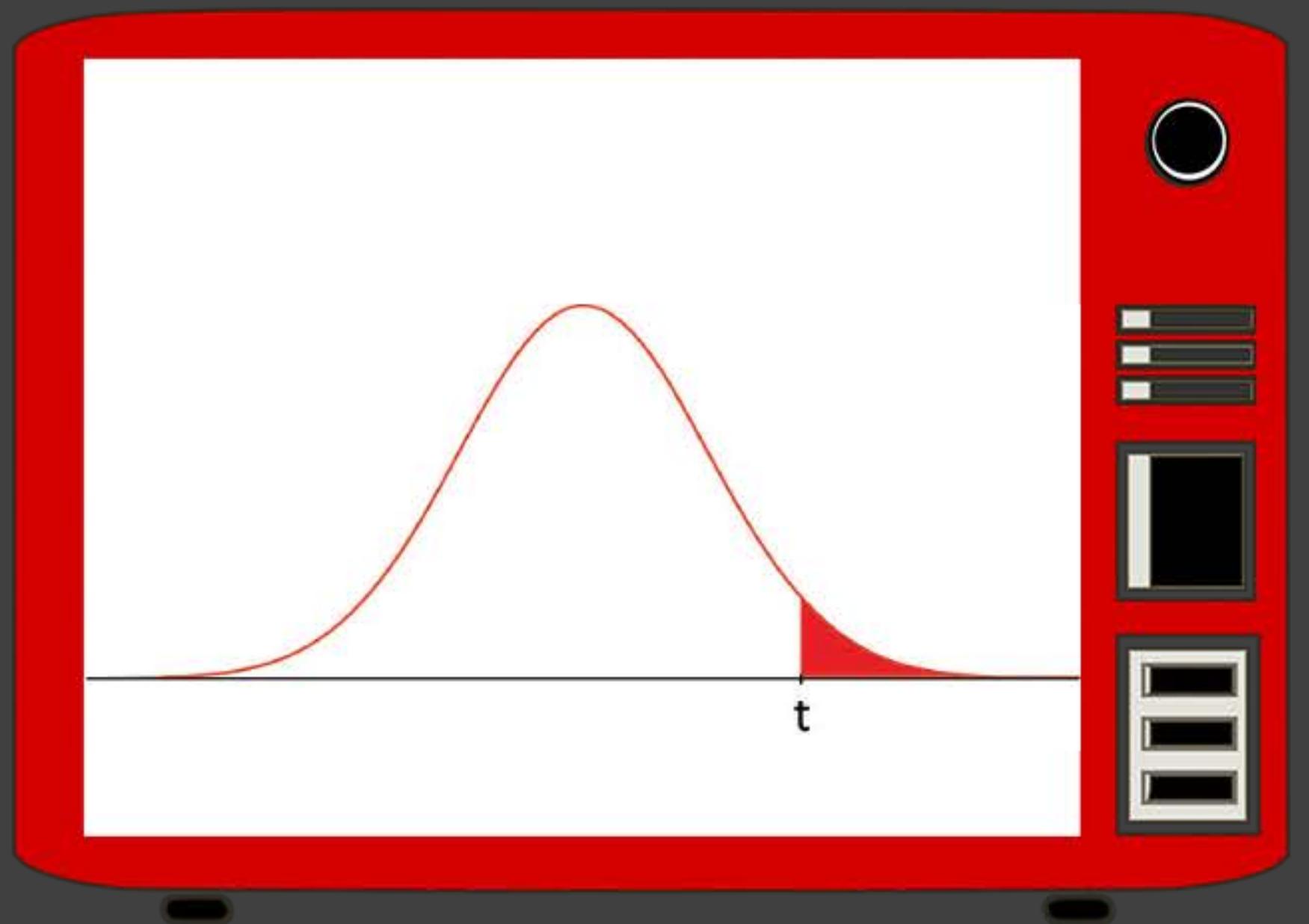
НУЛЕВОЕ РАСПРЕДЕЛЕНИЕ

- › $F(x)$ — нулевое распределение статистики
- › t — значение статистики на полученных данных.
- › Каким значениям статистики соответствует H_1 ?



НУЛЕВОЕ РАСПРЕДЕЛЕНИЕ

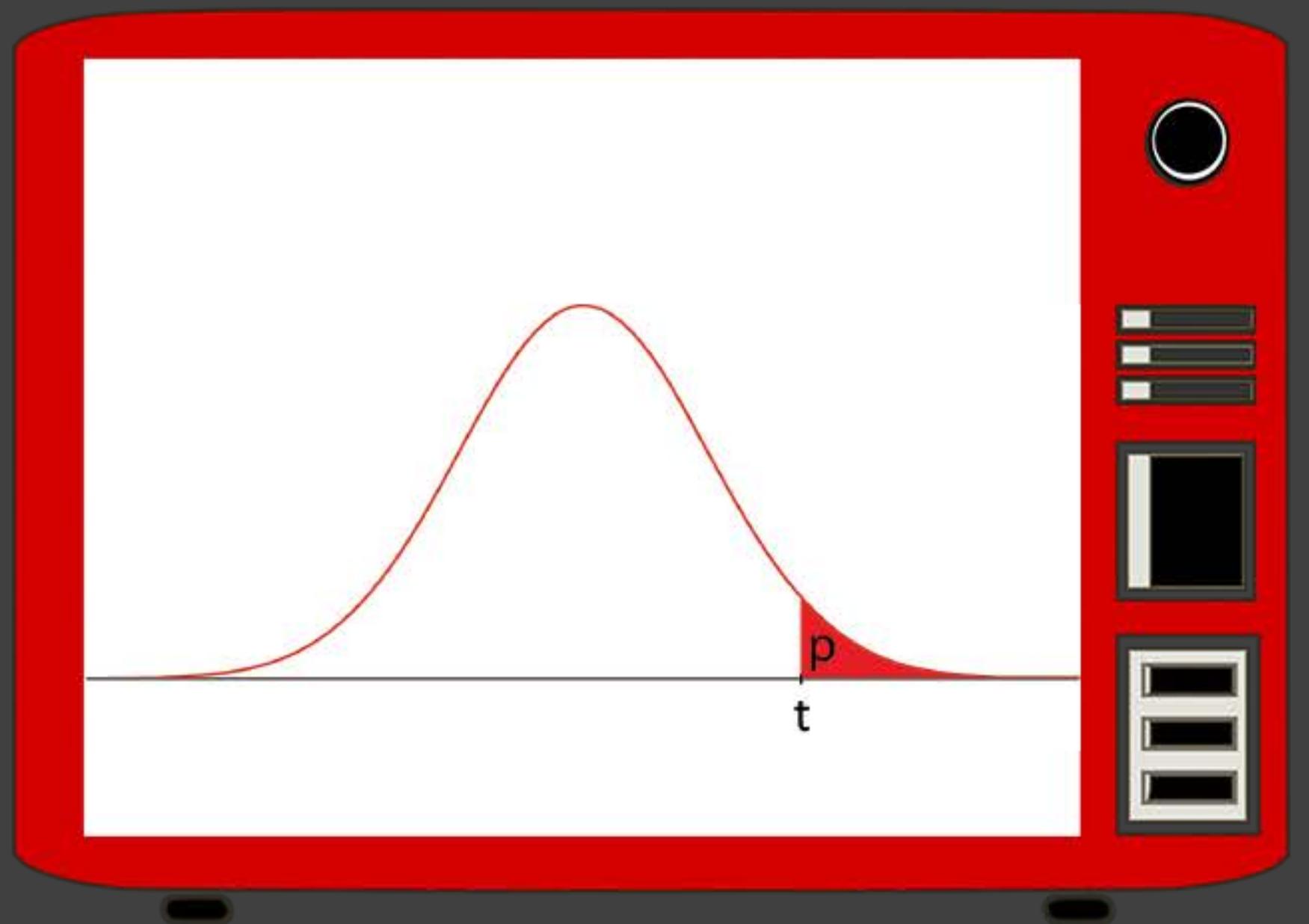
- » t — значение статистики на полученных данных.
- » Каким значениям статистики соответствует H_1 ?
- » Допустим, большим:



ДОСТИГАЕМЫЙ УРОВЕНЬ ЗНАЧИМОСТИ

- › Какова вероятность при H_0 получить значение t или больше?
- › Достигаемый уровень значимости (p-value):

$$p = \text{P}(T \geq t | H_0)$$



ДОСТИГАЕМЫЙ УРОВЕНЬ ЗНАЧИМОСТИ

- » Достигаемый уровень значимости (p-value):

$$p = \mathbf{P}(T \geq t | H_0)$$

- » p — вероятность при справедливости нулевой гипотезы получить значение статистики как в эксперименте или ещё более экстремальное
- » p мало \Rightarrow данные свидетельствуют против нулевой гипотезы в пользу альтернативы

ДОСТИГАЕМЫЙ УРОВЕНЬ ЗНАЧИМОСТИ

- › p — вероятность при справедливости нулевой гипотезы получить значение статистики как в эксперименте или ещё более экстремальное
- › p мало \Rightarrow данные свидетельствуют против нулевой гипотезы в пользу альтернативы
- › α — уровень значимости; H_0 отвергается в пользу H_1 при $p \leq \alpha$

➤ Механизм проверки гипотез:

- ▶ Гипотеза и альтернатива
- ▶ Статистика
- ▶ Нулевое распределение
- ▶ Достигаемый уровень значимости

- » Механизм проверки гипотез:
 - ▶ Гипотеза и альтернатива
 - ▶ Статистика
 - ▶ Нулевое распределение
 - ▶ Достигаемый уровень значимости

- » Далее: ошибки первого и второго рода

ОШИБКИ I И II РОДА

ОШИБКИ I И II РОДА

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка II рода
H_0 отвергается	Ошибка I рода	H_0 верно отвергнута

➤ Ошибки I и II рода не равнозначны!

ЗАДАЧА НЕСИММЕТРИЧНА

» Ошибка первого рода критичнее:

- $P(\text{отвергаем } H_0 | H_0)$ жёстко ограничивается:
если H_0 отвергается при $p \leq \alpha$, то вероятность ошибки первого рода:

$$P(H_0 \text{ отвергнута} \mid H_0 \text{ верна}) = P(p \leq \alpha \mid H_0) \leq \alpha$$

ЗАДАЧА НЕСИММЕТРИЧНА

- Ошибка первого рода критичнее:
 - ▶ $P(\text{отвергаем } H_0 | H_0)$ жёстко ограничивается
 - ▶ $P(\text{принимаем } H_0 | H_1)$ мягко минимизируется

Мощность критерия:

$$\text{pow} = P(\text{отвергаем } H_0 | H_1) = 1 - P(\text{принимаем } H_0 | H_1)$$

- Идеальный критерий имеет максимальную мощность

ЗАДАЧА НЕСИММЕТРИЧНА

- › H_0 и H_1 не равнозначны!
- › Нельзя доказать, что H_0 верна:
 - $p \leq \alpha \Rightarrow H_0$ отвергается в пользу H_1
 - $p > \alpha \Rightarrow H_0$ не отвергается в пользу H_1

ЗАДАЧА НЕСИММЕТРИЧНА

- › H_0 и H_1 не равнозначны!
- › Нельзя доказать, что H_0 верна:
 - ▶ $p \leq \alpha \Rightarrow H_0$ отвергается в пользу H_1
 - ▶ $p > \alpha \Rightarrow H_0$ не отвергается в пользу H_1
- › Отсутствие доказательств чего-то не является доказательством обратного!



- » Отличия между гипотезой и альтернативой
- » Ошибки I и II рода
- » Далее: достигаемый уровень значимости

ДОСТИГАЕМЫЙ УРОВЕНЬ ЗНАЧИМОСТИ

ДОСТИГАЕМЫЙ УРОВЕНЬ ЗНАЧИМОСТИ

$$p = \mathbf{P}(T \geq t | H_0)$$

- › Вероятность получить значение статистики как в эксперименте или ещё более экстремальное при справедливости нулевой гипотезы

ДОСТИГАЕМЫЙ УРОВЕНЬ ЗНАЧИМОСТИ

$$p = \mathbf{P}(T \geq t | H_0)$$

- › Вероятность получить значение статистики как в эксперименте или ещё более экстремальное при справедливости нулевой гипотезы
- › Чем ниже p , тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы

НЕПРАВИЛЬНАЯ ИНТЕРПРЕТАЦИЯ

- » $p = \mathbf{P}(T \geq t | H_0) \neq \mathbf{P}(H_0)$
 $\neq \mathbf{P}(H_0 | T \geq t)$

НЕПРАВИЛЬНАЯ ИНТЕРПРЕТАЦИЯ

- » $p = \mathbf{P}(T \geq t | H_0) \neq \mathbf{P}(H_0)$
 $\neq \mathbf{P}(H_0 | T \geq t)$
- » Осьминог угадал результаты 11 из 13 матчей с участием сборной Германии на чемпионате мира по футболу 2010 г.



НЕПРАВИЛЬНАЯ ИНТЕРПРЕТАЦИЯ

- » $p = \text{P}(T \geq t | H_0) \neq \text{P}(H_0)$
 $\neq \text{P}(H_0 | T \geq t)$
- » Осьминог угадал результаты 11 из 13 матчей
- » $p = 0.0112$ — не вероятность того, что осьминог выбирает кормушку наугад. Эта вероятность равна единице

- › Определение достигаемого уровня значимости не упрощаемо
- › Далее: статистическая и практическая значимость

СТАТИСТИЧЕСКАЯ И ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ

СТАТИСТИЧЕСКАЯ И ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ



- » Интерес представляет не p , а размер эффекта — степень отклонения данных от нулевой гипотезы

РАЗМЕР ЭФФЕКТА

- › Вероятность верного предсказания
- › Вероятность выздоровления пациента, принимавшего лекарство, минус вероятность выздоровления пациента, принимавшего плацебо
- › Увеличение среднего чека интернет-магазина при подключении программы лояльности

РАЗМЕР ЭФФЕКТА

- › Оценка размера эффекта по выборке — случайная величина
- › p показывает, с какой вероятностью такую оценку можно было получить случайно

p ЗАВИСИТ ОТ *n*

- › *p* зависит не только от размера эффекта, но и от размера выборки
- › По мере увеличения *n* H_0 может сначала приниматься, но потом выявляются более тонкие несоответствия выборки гипотезе H_0 , и она будет отвергнута

СТАТИСТИЧЕСКИ ЗНАЧИМО ПРАКТИЧЕСКИ НЕЗНАЧИМО



› Влияния физических упражнений на набор веса:

- ▶ за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$)

СТАТИСТИЧЕСКИ ЗНАЧИМО ПРАКТИЧЕСКИ НЕЗНАЧИМО



- Влияния физических упражнений на набор веса:
 - ▶ за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$)
- Разница в набранном весе составила 150 г. Практическая значимость такого эффекта сомнительна

СТАТИСТИЧЕСКИ ЗНАЧИМО ПРАКТИЧЕСКИ ЗНАЧИМО



- Гормональный препарат “Премарин”, облегчающий симптомы менопаузы:
- ▶ В 2002 году клинические испытания были досрочно прерваны. Было обнаружено, что приём ведёт к значительному увеличению риска развития рака груди (на 0.08%), риска инсульта (на 0.08%) и инфаркта (на 0.07%)

СТАТИСТИЧЕСКИ ЗНАЧИМО ПРАКТИЧЕСКИ ЗНАЧИМО



- » В 2002 году клинические испытания были досрочно прерваны. Было обнаружено, что приём ведёт к значительному увеличению риска развития рака груди (на 0.08%), риска инсульта (на 0.08%) и инфаркта (на 0.07%)
- » Эффект выглядит маленьким, но с учётом численности населения он превращается в тысячи дополнительных смертей

СТАТИСТИЧЕСКИ НЕЗНАЧИМО ПРАКТИЧЕСКИ ЗНАЧИМО



- Лекарство, замедляющее ослабление интеллекта больных Альцгеймером:
 - ▶ При испытаниях оказывается, что разница в IQ контрольной и тестовой групп составляет 13 пунктов, но разница статистически незначима
- Возможно, изучение лекарства следует продолжить

- › Всегда оценивайте размер эффекта!
- › Статистическая и практическая значимость
- › Далее: первые примеры статистических критериев

ПРОВЕРКА ГИПОТЕЗ И ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

КАЧЕСТВО КЛАССИФИКАТОРА

- › Бинарный классификатор правильно предсказывает метку класса на **60** из **100** объектах тестовой выборки
- › Можно ли считать, что он лучше, чем генератор случайных чисел?

КАЧЕСТВО КЛАССИФИКАТОРА

- › Бинарный классификатор правильно предсказывает метку класса на 60 из 100 объектах тестовой выборки
- › 95% нормальный доверительный интервал для доли верно предсказанных меток: [0.504, 0.696]
- › Не содержит 0.5!

ИНТЕРВАЛ \Rightarrow ГИПОТЕЗА



$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

ИНТЕРВАЛ \Rightarrow ГИПОТЕЗА

- » $H_0: \theta = \theta_0$
- $H_1: \theta \neq \theta_0$
- ↑
- » H_0 отвергается на уровне значимости α , если $100(1 - \alpha)\%$ интервал для θ не содержит θ_0

ИНТЕРВАЛ \Rightarrow ГИПОТЕЗА

- » $H_0: \theta = \theta_0$
- $H_1: \theta \neq \theta_0$
- ↑
- » H_0 отвергается на уровне значимости α , если $100(1 - \alpha)\%$ интервал для θ не содержит θ_0

- » Достигаемый уровень значимости p — наибольшее α , при котором $100(1 - \alpha)\%$ доверительный интервал содержит θ_0

КАЧЕСТВО КЛАССИФИКАТОРА

- › Бинарный классификатор правильно предсказывает метку класса на 60 из 100 объектах тестовой выборки
- › 95% нормальный доверительный интервал Уилсона для доли верно предсказанных меток: [0.502, 0.691]
- › Соответствующий достигаемый уровень значимости: $p = 0.045$

ГИПОТЕЗА \Rightarrow ИНТЕРВАЛ

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

ГИПОТЕЗА \Rightarrow ИНТЕРВАЛ

- » $H_0: \theta = \theta_0$
- $H_1: \theta \neq \theta_0$
- ↓
- » $100(1 - \alpha)\%$ доверительный интервал для θ состоит из всех значений θ_0 , для которых H_0 не отвергается на уровне значимости α

КАЧЕСТВО КЛАССИФИКАТОРОВ

- » Первый классификатор правильно предсказывает метку класса на **60** из **100**, второй — на **75** из **100** объектах тестовой выборки

КАЧЕСТВО КЛАССИФИКАТОРОВ

- › Первый классификатор правильно предсказывает метку класса на **60 из 100**, второй — на **75 из 100** объектах тестовой выборки
- › Можно ли считать, что второй классификатор лучше?

КАЧЕСТВО КЛАССИФИКАТОРОВ

- › Первый классификатор правильно предсказывает метку класса на **60 из 100**, второй — на **75 из 100** объектах тестовой выборки
- › **95%** доверительный интервал Уилсона для доли верных предсказаний первого классификатора: **[0.502, 0.691]**. Второго: **[0.657, 0.825]**

КАЧЕСТВО КЛАССИФИКАТОРОВ

- » 95% доверительный интервал Уилсона для доли верных предсказаний первого классификатора: [0.502, 0.691]. Второго: [0.657, 0.825]
- » Доверительные интервалы пересекаются, значит, классификатора неотличимы по качеству?

ПАРА ИНТЕРВАЛОВ $\not\Rightarrow$ ГИПОТЕЗА

- › $H_0: \theta_1 = \theta_2$
 $H_1: \theta_1 \neq \theta_2$
- › Если доверительные интервалы для θ_1 и θ_2 пересекаются, это ещё не значит, что H_0 нельзя отвергнуть!

КАЧЕСТВО КЛАССИФИКАТОРОВ

- » 95% доверительный интервал для разности долей:
[0.022, 0.278]
- » Соответствующий достигаемый уровень значимости: $p = 0.022$

КАЧЕСТВО КЛАССИФИКАТОРОВ

- » В задаче выборки связанные:

I	II	+	-	Σ
+	55	5	60	
-	20	20	40	
Σ	75	25	100	

КАЧЕСТВО КЛАССИФИКАТОРОВ

- » 95% доверительный интервал для разности долей в связанных выборках: [0.06, 0.243]
- » Соответствующий достигаемый уровень значимости: $p = 0.002$

➤ Проверка гипотез:

- ▶ Устройство статистических критериев
- ▶ Достигаемый уровень значимости
- ▶ Размер эффекта
- ▶ Связь с доверительными интервалами