Проверка статистических гипотез • Hypothesis testing

Сегодня

Эльбрус буткемп

- посмотрим, как отличить статистически достоверное событие от случайного
- узнаем как устроен пайплайн проверки гипотез
- реализуем функции расчета для некоторых тестов

Проверка статистических гипотез

Эльбрус буткемп

- Статистическая гипотеза – предположение о свойствах генеральной совокупности.
- Всю ГС мы исследовать не можем, значит, мы должны собрать **репрезентативную выборку**, изучить ее, а после проверить гипотезу.

Задача с кофе

Эльбрус буткемп

Задача с кофе

Эльбрус буткемп

- в ТЦ стоит один из наших автоматов с кофе.
- **Ранее**:
  - из тех, кто подходил к нему, с кофе уходил каждый второй.
- **Есть гипотеза**:
  - У нашего автомата сложный интерфейс, и некоторых людей это сбивает с толку и они уходят
- **Изменения**:
  - Разработан новый тестовый интерфейс, и поставлен на наш автомат.
  - В случае успеха, можно будет выкатывать на остальные наши автоматы
- **После** :
  - Из 300 людей, которые подошли к нашему автомату, купили 167

Задача

Эльбрус буткемп

- Подтвердилась ли наша гипотеза? К какому результату интуитивно склоняетесь?
- Желательно не ошибиться с выбором, так как внедрение нового интерфейса на все наши автоматы стоит денег.

Задача

Эльбрус буткемп

Формализуем:

- **ГС** - все люди, которые подошли бы к нашему автомату
- у нас есть **выборка** $x_1, x_2, x_3, \ldots, x_{300}$
- $x_i \sim Be(p)$ (*купил/не купил*)
- p - неизвестный для нас параметр ГС - доля тех, кто купил бы, подойдя к автомату
- Пример выборки $[1, 1, \ldots, 1, 0, 1, 0, 1, 1]$ (*167 купили, 133 не купили*)

Статистическая гипотеза

Эльбрус буткемп

Нулевая гипотеза H_0 – это гипотеза, которой мы придерживаемся, пока наблюдения не заставят признать обратное. Ей всегда сопутствует альтернативная гипотеза H_1.

- H_0 почти всегда формулируется, как "значимых изменения нет"
- H_1 - "значимые изменения есть"

Статистическая гипотеза

Эльбрус буткемп

В нашем случае:

- $H_0 : p = 0.5$ (конверсия в покупку такая же и осталась)
- $H_1 : p > 0.5$ (конверсия увеличилась)

По результатам исследования мы остановимся на одной из гипотез

Ошибка первого и второго рода

Эльбрус буткемп

- Ошибка первого рода(FP) - это ситуация, когда H_0 отвергается, хотя она, на самом деле, верна
  - α – вероятность ошибки первого рода или уровень значимости
- Ошибка второго рода(FN) - это ситуация, когда H_0 принимается, хотя она неверна
  - β – вероятность ошибки второго рода

Некоторая сложность: при уменьшении ошибки первого рода, увеличивается ошибка второго рода и наоборот.

Ошибка первого и второго рода

- Матрица ошибок (confusion matrix)

Hypothesis testing:

Actual \ Decision	H_0 true (Fail to reject)	H_0 false (Rejecting H_0)
H_0 true	TRUE NEGATIVE Correct decision: **Confidence level** (prob $1-\alpha$)	FALSE POSITIVE **Type I Error**: **Significance level/Size (α)** (prob α)
H_0 false	FALSE NEGATIVE **Type II Error**: fail to reject (prob β)	TRUE POSITIVE Correct decision: **Power** (prob $1-\beta$)

Примеры

Эльбрус буткемп

- Например, тест на COVID-19 показывает отрицательный результат у пациента, который на самом деле инфицирован. Ошибка какого рода?
- "Ложноположительный результат" при анализе маркетинговой кампании – это ошибка какого рода?

Статистика критерия

Эльбрус буткемп

Статистика – любая функция, получаемая по выборке. В каком-то смысле это просто посчитанная метрика

Обозначение: $\mathrm{T}(\vec{x})$, где $\vec{x} = (x_1, x_2, x_3, \ldots, x_n)$ - выборка

- Статистика агрегирует информацию о выборке.
- Самые частые статистики:
  - Среднее, доля, медиана, количество, квантиль и т.д
- Кастомные статистики
  - Показатель удовлетворенности клиентов (Customer Satisfaction Score, CSS)
  - Показатель устойчивости бизнеса(Customer Loyalty Index, CLI)
  - "Здесь бы могла быть ваша статистика🙂"

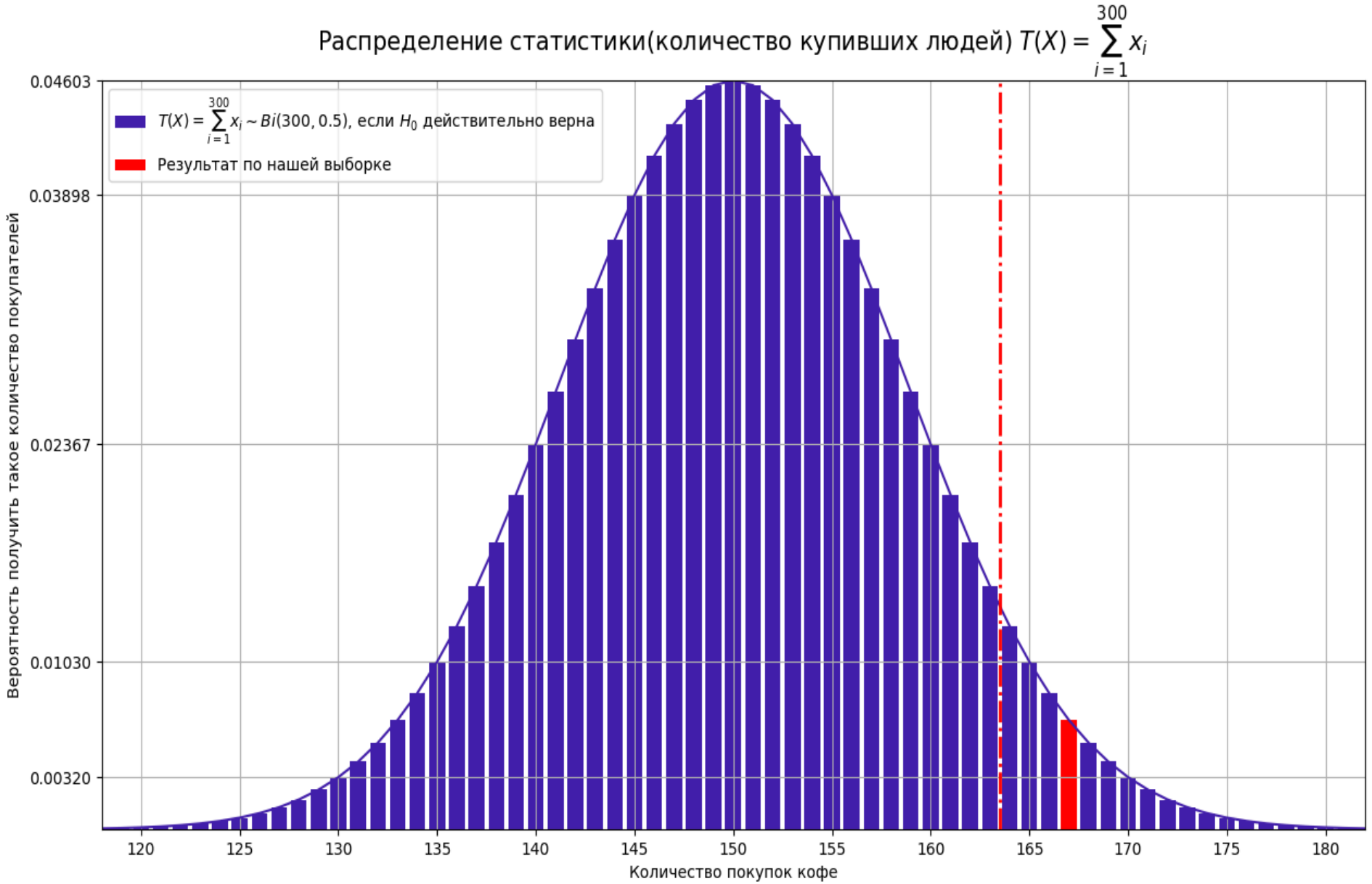
Наша задача

Эльбрус буткемп

- Возьмем в нашем примере с кофе $T(X) = \sum_{i=1}^{300} x_i$, иначе говоря, сколько людей купили у нас кофе.
- Важно понимать, что $T(X)$ тоже является **случайной величиной**, а значит имеет свое распределение, это **ключевой момент** в данной теме.
- Именно знание распределения статистики дает нам понимания, насколько **экстремальное** значение мы вообще получили.
- Например при проверки монетки на честность получить 90 орлов после 100 подбрасываний кажется слишком экстремальным, и скорее она не честная.

Наша задача

Эльбрус буткемп



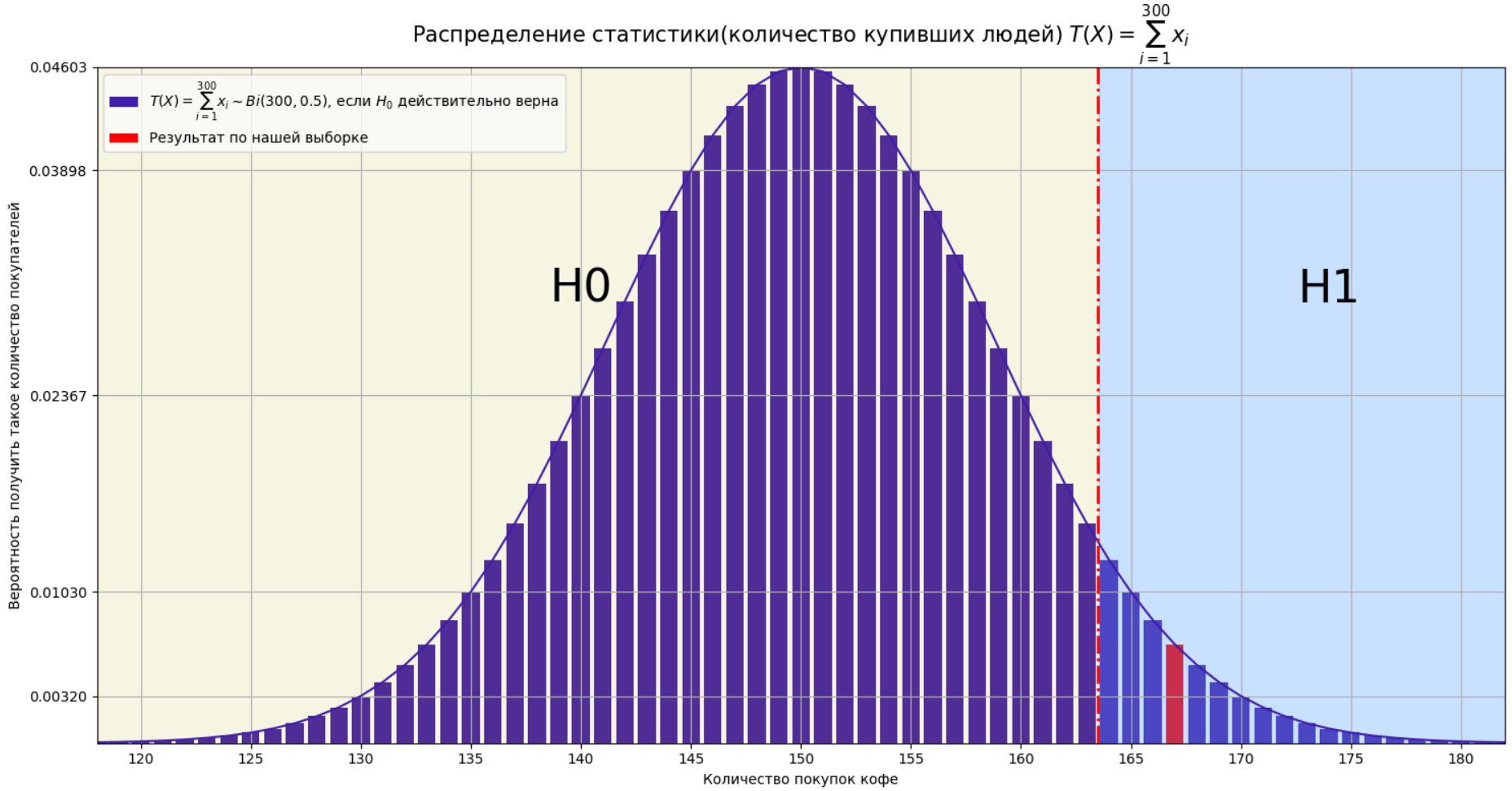
Проверка гипотез: алгоритм

Эльбрус буткемп

1. Сформулировать основную и альтернативную гипотезы, задать уровень значимости α
2. Найти критические значения статистики для соответствующего уровня значимости
3. Вычислить значение статистики и определить, попало ли оно в критическую область
4. Сделать вывод: если значение попало в критическую область - отвергнуть нулевую гипотезу, в противном случае принять

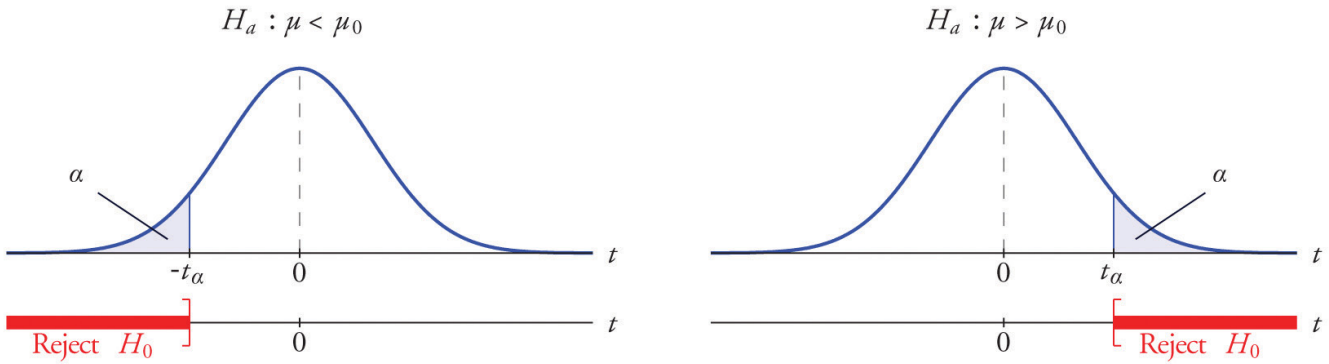
Наша задача

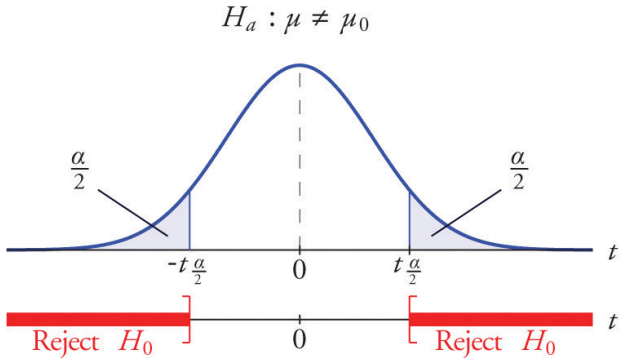
- α = 0.05 - уровень значимости или ошибка первого рода



Критическая область

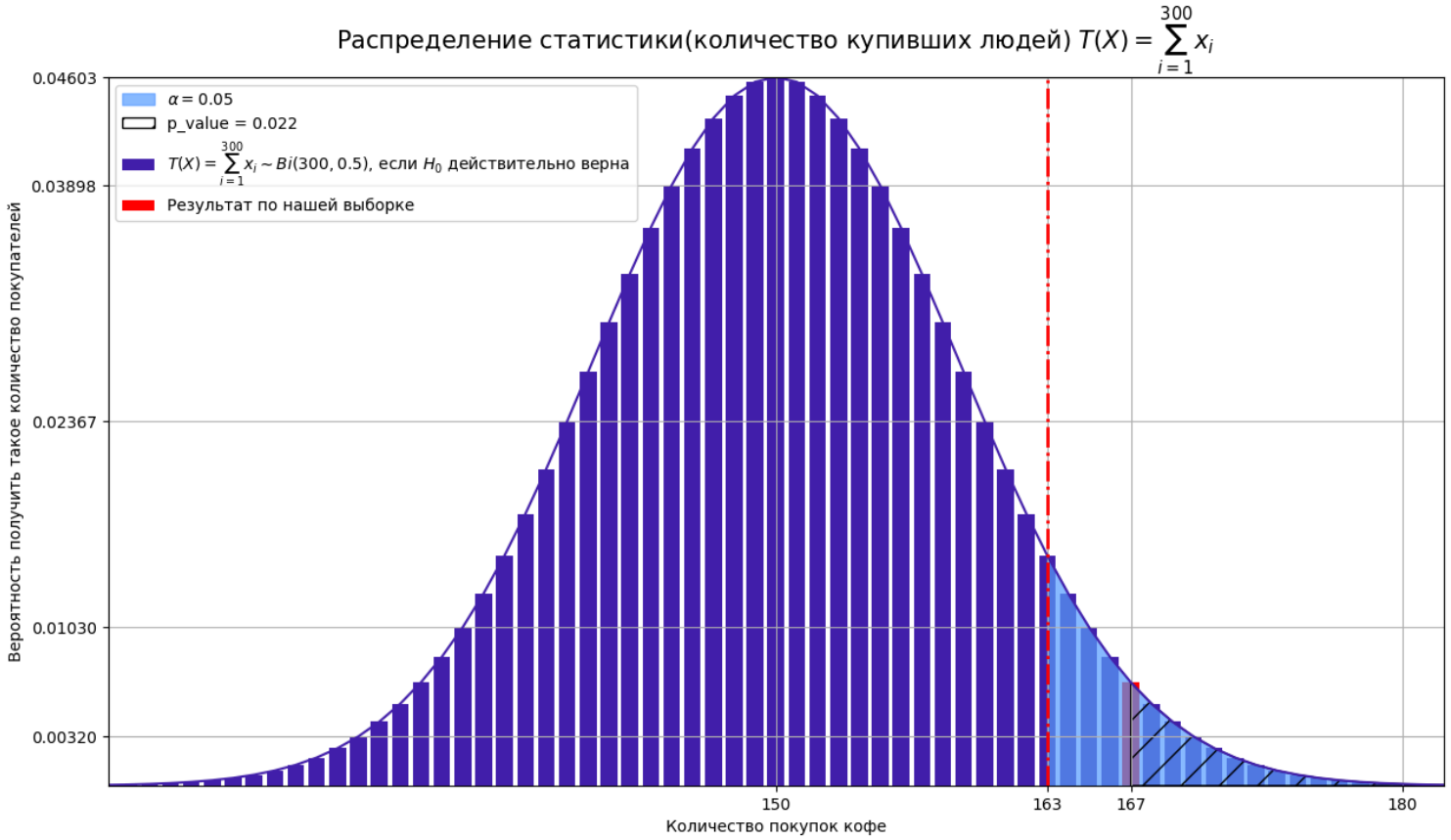
Критической областью называется область значений статистики критерия, при которых отвергается H_0. А критические значения - это граница критической области.





p-value

- **p-value** можно интерпретировать как вероятность ошибиться, если мы выбираем гипотзу H1.



Основные статистические тесты

Проверка гипотезы о среднем

Эльбрус буткемп

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

> Среднее генеральной совокупности μ равно некоторому предполагаемому среднему μ_0

Проверка гипотезы о среднем

Эльбрус буткемп

В случае **известного стандартного отклонения** для всей ГС, статистика критерия

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim \mathbb{N}(0, 1)$$

имеет нормальное распределение в предположении справедливости гипотезы H_0.

Зная вид распределения у статистики z, мы можем оценить, насколько вероятно было получить посчитанное нами значение и на основе этого делать выводы!

- Для левосторонней области необходимо найти квантиль уровня α
- Для правосторонней уровня $1 - \alpha$
- Для двусторонней – квантили уровней $\frac{a}{2}$ и $1 - \frac{a}{2}$

> Если стандартное отклонение известно, то используется z-критерий.

Проверка гипотезы о среднем: пример

Эльбрус буткемп

Предположим, что мы хотим проверить, действительно ли средний вес яблок в саду составляет 150 грамм. Мы собрали выборку из нескольких яблок и получили следующие данные: $145, 152, 148, 151, 149, 153$. Известно, что стандартное отклонение веса яблок $\sigma = 2$. Проверим гипотезу, что средний вес больше 150 грамм, на уровне значимости $\alpha = 0.05$.

Проверка гипотезы о среднем: пример

Эльбрус буткемп

1. $H_0 : \mu = 150,\ H_1 : \mu > 150$
2. Критическая область правосторонняя: $1 - \alpha = 0.95$, критическое значение $z_{cr} = 1.645$. Чтобы отвергнуть H_0, z должно быть больше 1.645.
3. $\bar{x} = \dfrac{145 + 152 + 148 + 151 + 149 + 153}{6} = 149.67,\ \sigma = 2,\ n = 6$. Тогда
$$z = \frac{149.67 - 150}{2/\sqrt{6}} \approx -0.68$$
4. Т.к. $z < z_{cr}$, нулевая гипотеза не отвергается.

Проверка гипотезы о среднем

Эльбрус буткемп

Если стандартное отклонение гипотезы неизвестно или же имеем небольшую выборку, то используется t-статистика (aka тест Стьюдента, t-тест).

Процедура остается прежней, но статистика меняется:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Тогда $t \sim St(n-1)$ - Распределение Стьюдента

s - выборочная оценка стандартного отклонения:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=i}^{n}(x_i - \bar{x})^2}$$

Проверка гипотезы о среднем: пример

Эльбрус буткемп

Для сравнения двух выборок между собой, t-тест применим тоже.

$H_0: \mu 1 = \mu 2$

$H_1: \mu 1 \neq \mu 2$

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Тогда $t \sim St(n_1 + n_2 - 2)$ - Распределение Стьюдента

μ_i, s_i^2, n_i - параметры i-ой выборки

Проверка гипотезы о среднем: пример

Эльбрус буткемп

- Дана выборка: $1, 0, 3, 5, 4$, уровень значимости $\alpha = 0.01$, гипотезы:

$$H_0 : \mu = 3,\ H_1 : \mu < 3$$

- Критическая область левосторонняя, число степеней свободы $df = n - 1 = 4$, по квантилю уровня α определяем критическое значение $t_{cr} = -3.74$
- Вычисляем значение статистики: $t_{st} = 0.64$
- Вывод: $t_{st} \notin (-\infty; -3.74)$ => H_0 не отвергается.

Гипотезы о виде распределения

Эльбрус буткемп

- **U-критерий Манна — Уитни**, который сравнивает два распределения. Данный тест в отличие от многих других является **непараметрическим**
- H_0 : две группы имеют одинаковое распределение
- H_1 : одна из групп имеет большие (или меньшие) значения, чем другая

$U_1 = R_1 - n_1(n_1 + 1)/2$, где R_1 — сумма рангов точек данных в первой группе, а n_1 — количество точек в первой группе.

$$U_2 = R_2 - n_2(n_2 + 1)/2$$

$$T_{stat} = min(U_1, U_2) \sim N(\frac{n_1 * n_2}{2}, \frac{n_1 * n_2 * (n_1 + n_2 + 1)}{12})$$

Гипотеза о доле

Эльбрус буткемп

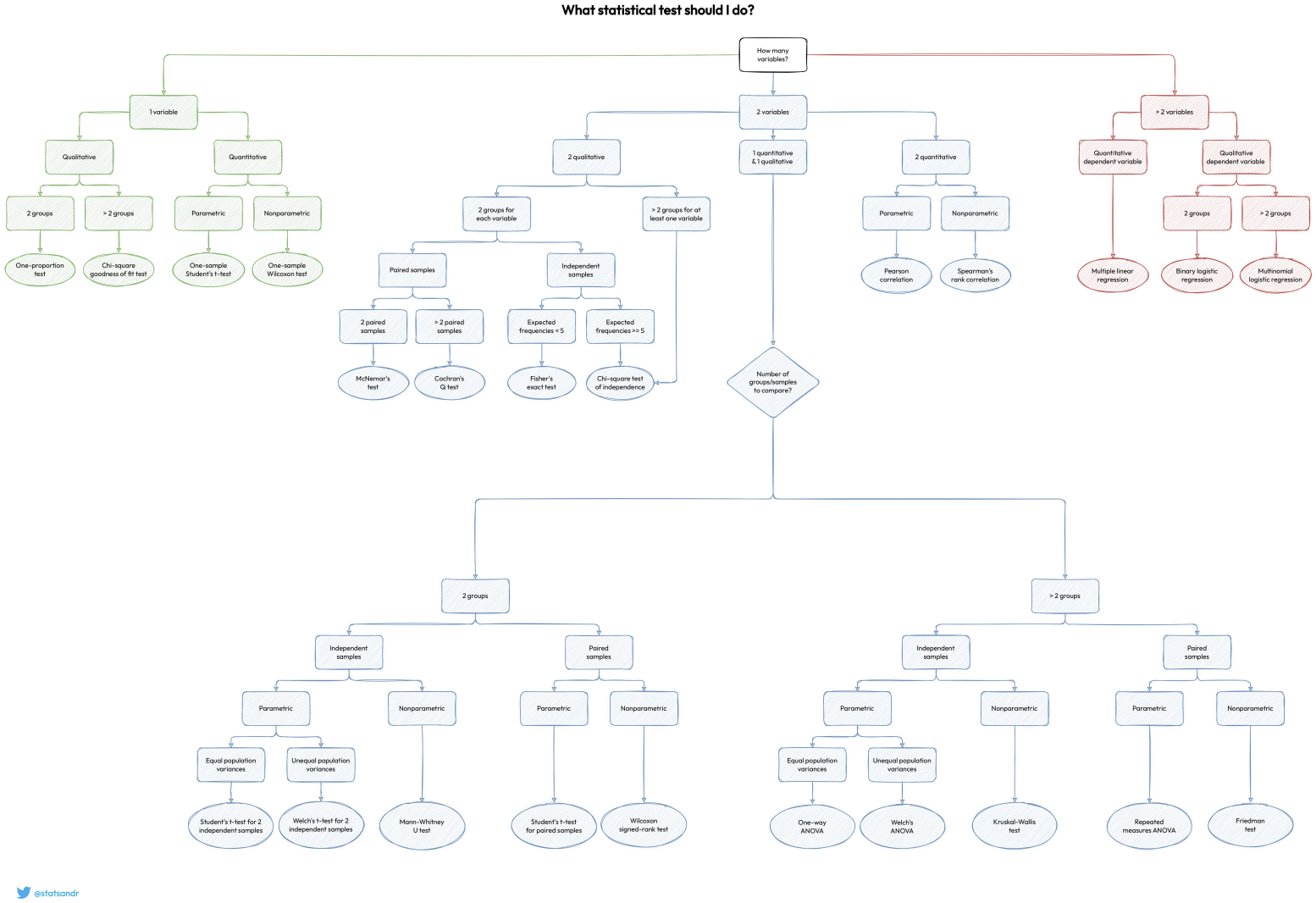
- Проверяем гипотезу $H_0 : p = p_0$
- Значение статистики вычисляется как

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

n - объем выборки, p_0 предполагаемая доля носителей признака, $\hat{p}$ - выборочная доля носителей признака.

> Проверим гипотезу, что доля признака в ГС равна 0.1 на уровне значимости $\alpha = 0.05$, против односторонних альтернатив $p > 0.1$.
> Объем выборки $n = 100$ и пусть выборочная доля составила $\hat{p} = 0.2$

Карта статистических тестов



Доверительные интервалы

Эльбрус буткемп

- Часто точечной оценки какой-либо метрики недостаточно.
- Можно спросить у двух людей их рост, вычислить среднее и получить какую-то величину, которая будет крайне мало связана с величиной, вычисленной по генеральной совокупности.
- Для этого можно построить доверительный интервал: интервал, который с заданной вероятностью накрывает истинное значение.

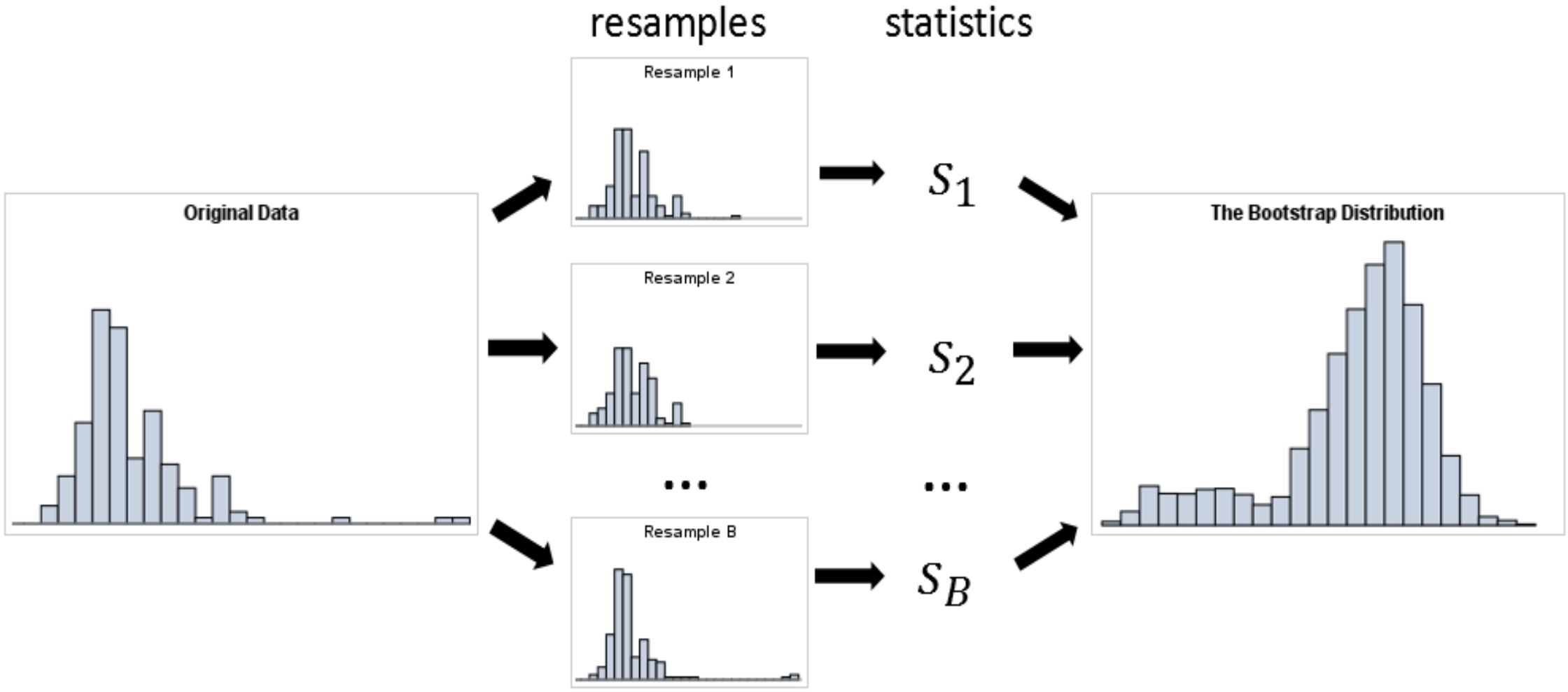
Доверительные интервалы

Эльбрус буткемп

- В самом общем случае, вместо квантиля стандратного нормального распределения нужно использовать квантиль распределения Стьюдента (t - распределения):

$$(\bar{x} - \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)); (\bar{x} + \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1))$$

Bootstrap



- Позволяет эмпирически получить доверительный интервал для измеряемой метрики

Bootstrap

Эльбрус буткемп

- Алгоритм построения бутстрап доверительного интервала:

1. Из исходной выборки размера n берем n случайных наблюдений с возвращением
2. Для полученной выборки вычисляем интересующую нас статистику
3. Повторяем шаги 1-2 много раз (например, 1000)
4. Получаем эмпирическое распределение статистики
5. Берем квантили этого распределения в качестве границ доверительного интервала

- Преимущества:
  - Не требует предположений о распределении данных
  - Работает для любых статистик(метрик)
  - Прост в реализации

Итоги

Эльбрус буткемп

- статистика применяется везде, где может применяться
- распределения - важная часть работы с данными
- визуализация очень важна
- проверка статистических гипотез - очень тонкая процедура
- Статистические тесты позволяют ответить есть ли **статистически значимый результат**
- Ошибки 1-го рода(α) и 2-го рода(β) не хороши для нас, однако в большинстве случаев их не избежать. Катастрофичность каждой из них зависит от конкретной задачи
- **p-value** позволяет оценить вероятность ошибки и принять решение