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

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)

Эльбрус буткемп

Hypothesis testing:

Actual	Decision: H_0 true (Fail to reject)	Decision: 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, суд выдвигает гипотезу H_0: подсудимый невиновен. А он, на самом деле, виновен, но суд признает его невиновным за отсутствием улик (презумпция невиновности). То есть суд принимает гипотезу, хотя она неверна.
- "Ложноположительный результат" при при медицинских анализах – 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🙂"

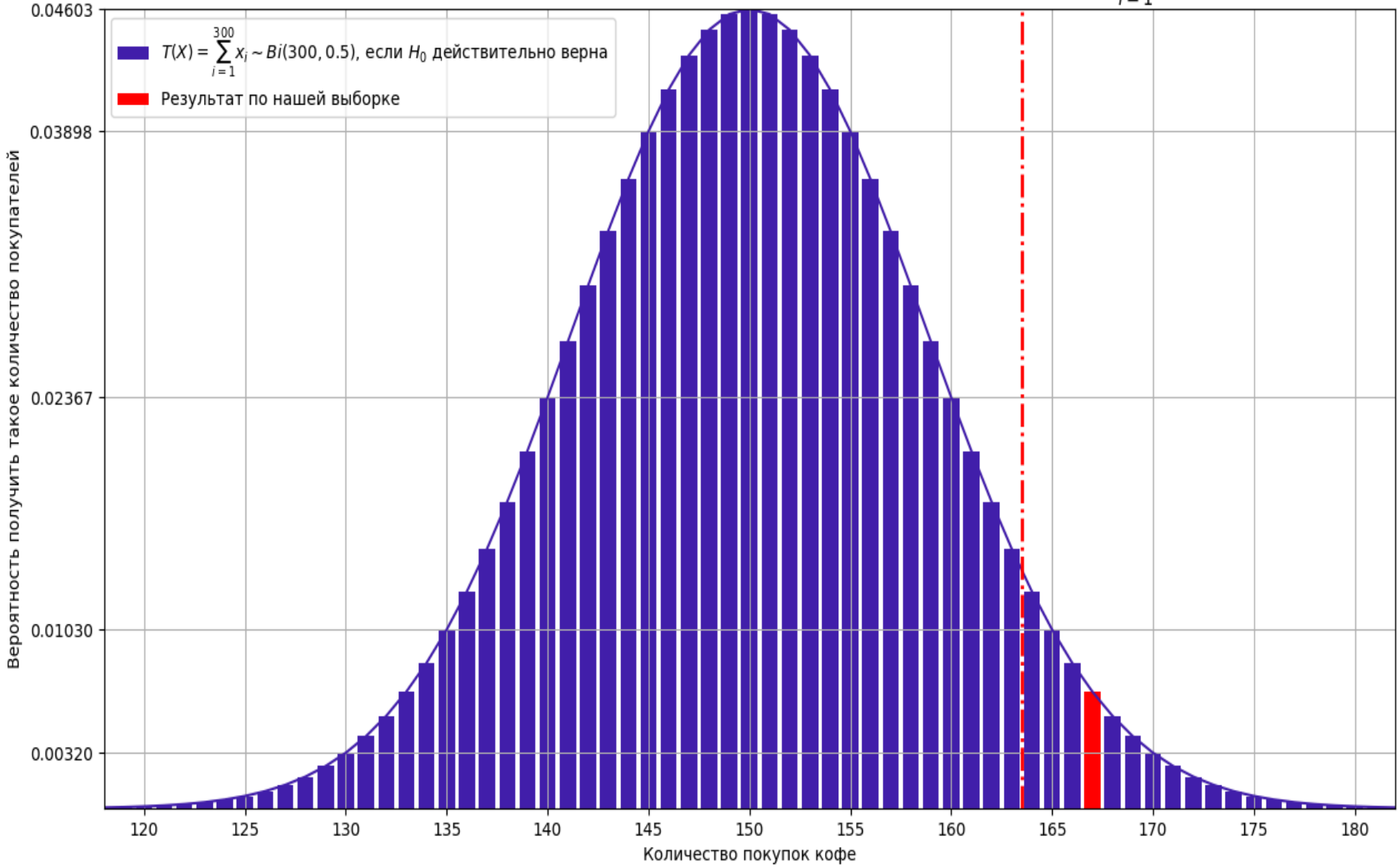
Наша задача

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

Распределение статистики(количество купивших людей) $T(X) = \sum_{i=1}^{300} x_i$



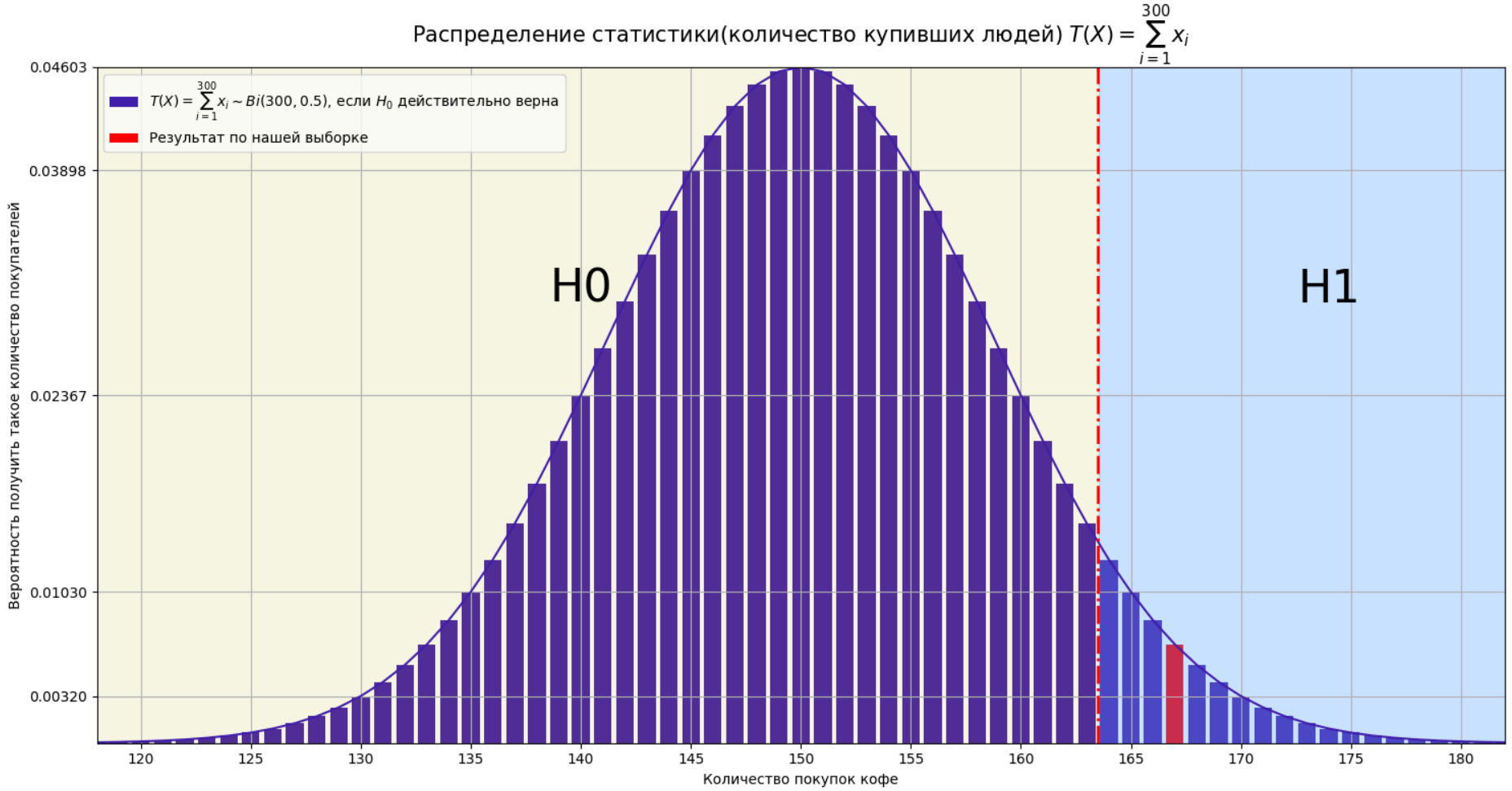
Проверка гипотез: алгоритм

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

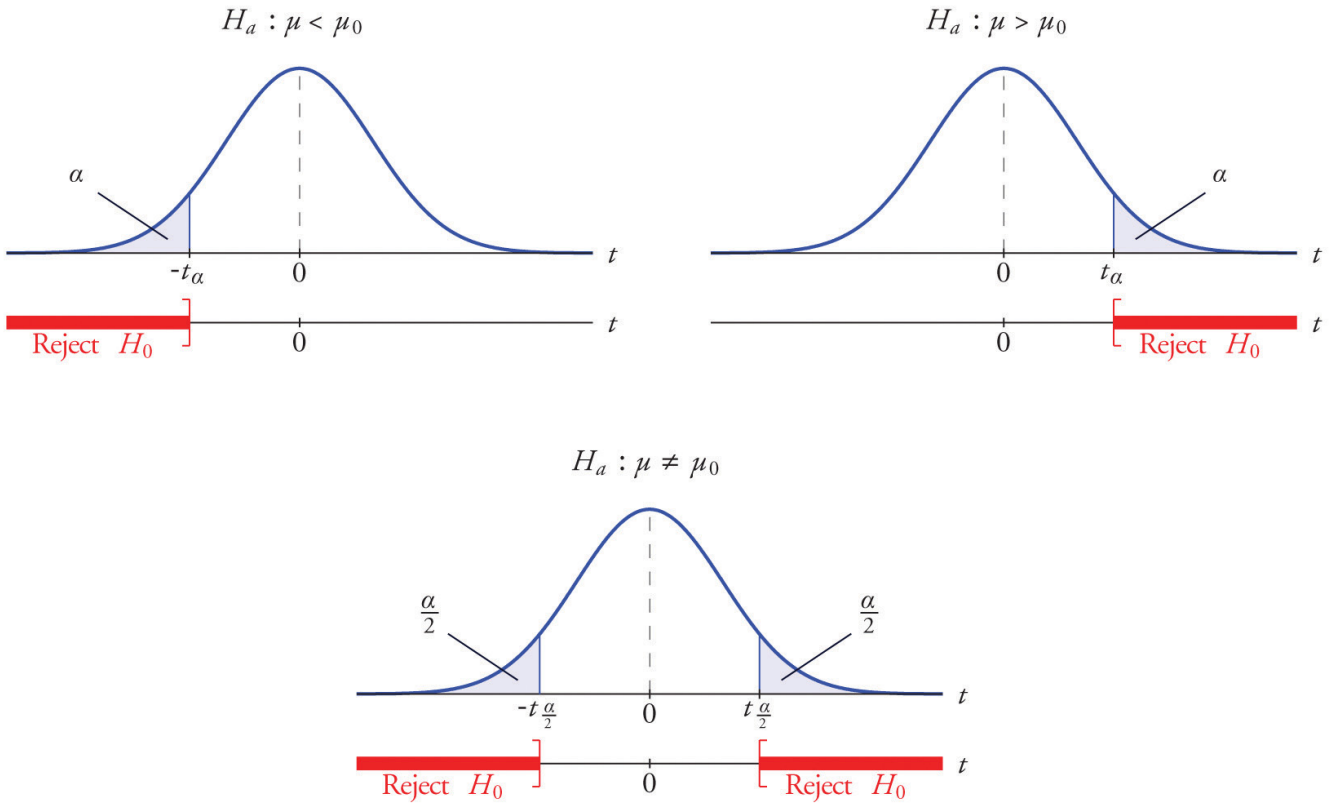
Наша задача

- α = 0.05 - уровень значимости или ошибка первого рода



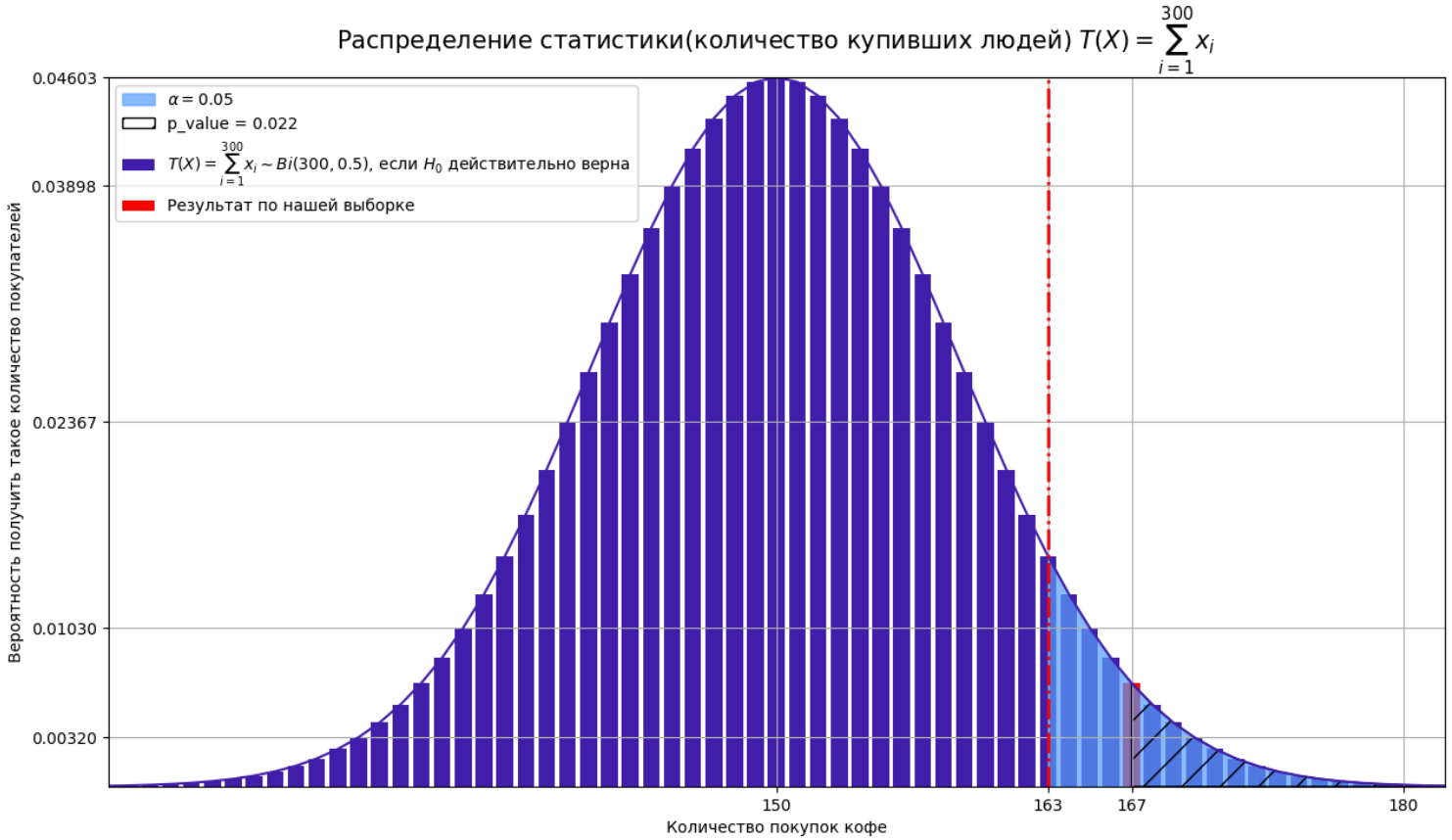
Критическая область

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.



Итоги

Эльбрус буткемп

- 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