Санкт-Петербургский государственный университет
Прикладная математика и информатика
Прикладная кибернетика

Научная иследовательская разбота

Сравние на синтетических данных и на данных с Kaggle методов для рассчёта доверительных интервалов для 90, 99 квантилей

Выполнил:
Кривоносов Тимофей Игоревич

Научный руководитель:
д. ф.-м. н., профессор
М. В. Юлдашев

Санкт-Петербург
2023

Оглавление

1. Введение

В современном мире статистический анализ играет важную роль во многих областях, от науки до бизнеса. При работе с данными мы часто сталкиваемся с необходимостью оценки параметров и построения доверительных интервалов для различных квантилей распределений. Квантили представляют собой значения, разделяющие вероятностное распределение на равные части.

В данной курсовой работе мы сосредоточимся на сравнении методов для расчета доверительных интервалов для 90-го и 99-го квантилей на основе синтетических данных и данных, предоставленных платформой Kaggle. Синтетические данные являются созданными искусственным образом наборами данных, которые позволяют нам контролировать различные аспекты, такие как распределение, объем выборки и т.д. Данные, полученные с Kaggle, представляют собой реальные данные, предоставленные сообществом пользователей этой платформы.

Для синтезирования данных и проверки методов поиска доверительных интервалов для 90 и 99 квантилей в задаче сравнения производительности и задержки (latency), проведу следующие шаги:

- Собрать данные: Запустите эксперименты или тесты, которые измеряют производительность и задержку вашей системы. Запишите результаты для каждого теста.
- Подготовить выборки: Создайте выборки из результатов тестов, соответствующие каждому методу или условию, которое вы хотите сравнить. Например, если вы сравниваете два разных алгоритма, создайте две выборки, содержащие результаты для каждого алгоритма.
- Анализировать выборки: Используйте статистические методы для анализа выборок и вычисления доверительных интервалов. Для нахождения доверительного интервала 90% квантили, найдите 5-й и 95-й перцентили выборки. Для доверительного интервала 99% квантили найдите 0.5-й и 99.5-й перцентили выборки.
- Проверьте методы поиска доверительных интервалов: Сравните результаты, полученные различными методами поиска доверительных интервалов. Убедитесь, что методы возвращают адекватные и интерпретируемые результаты.

- Интерпретация результатов: Проанализируйте полученные доверительные интервалы и сделайте выводы относительно производительности и задержки системы. Например, если интервалы значительно отличаются для двух методов, это может указывать на значимые различия между ними.

Основной целью нашего исследования является провести сравнительный анализ методов расчета доверительных интервалов для выбранных квантилей на основе обоих типов данных. Это позволит нам оценить, насколько точно и надежно эти методы работают в различных условиях и с разными типами данных.

Ожидается, что результаты данного исследования помогут нам лучше понять применимость и эффективность различных методов для расчета доверительных интервалов на основе разных типов данных, что может быть полезным при принятии статистических решений в реальных ситуациях.

2. Методы нахождения доверительных интервалов

Квантиль - это значение, которое разделяет упорядоченную выборку на две части таким образом, что заданный процент значений находится слева (ниже) этого значения, а оставшиеся значения находятся справа (выше). Например, 90-й квантиль (0.9-квантиль) разделяет выборку так, что 90% значений находятся ниже этого значения.

Доверительный интервал для заданного квантиля позволяет оценить диапазон, в котором с определенной вероятностью находится истинное значение параметра. Например, доверительный интервал для 90-го квантиля указывает на диапазон значений, в котором с вероятностью 90% будет находиться искомая величина.

2.1. Наивный подход

Наивный подход для поиска доверительного интервала для 90-го квантиля (или любого другого квантиля) является простым и интуитивным методом. Он основан на сортировке выборки значений и нахождении соответствующего элемента, который делит выборку на две части: одна содержит 90% значений, меньших или равных этому элементу, а другая - 10% значений, больших или равных ему.

Для нахождения доверительного интервала с использованием наивного подхода для 90-го квантиля, следуйте этим шагам:

- Упорядочите выборку значений в порядке возрастания (или убывания).
- Вычислите индекс элемента, который соответствует 90-му перцентилю, используя формулу: index = (90/100) * n, где n - размер выборки.
- Округлите индекс до ближайшего целого числа. Если индекс не является целым числом, возьмите следующее целое число больше него.
- Найдите значение, соответствующее округленному индексу в отсортированной выборке. Это будет значение, разделяющее выборку на 90% и 10% части.
- Доверительный интервал для 90-го квантиля будет состоять из значений, которые меньше или равны найденному значению.

Например, если у вас есть выборка [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] и вы хотите найти доверительный интервал для 90-го квантиля, вычисления будут следующими:

- Сортировка выборки: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10].
- Индекс = (90/100) * 10 = 9.
- Округленный индекс = 9.
- Значение по округленному индексу: 9.
- Доверительный интервал: [1, 2, 3, 4, 5, 6, 7, 8, 9].

Наивный подход для поиска доверительных интервалов может быть простым и быстрым, но он не учитывает особенности распределения данных и может быть менее точным в некоторых случаях. Поэтому для более точных и надежных результатов рекомендуется использовать статистические методы, такие как методы максимального правдоподобия или бутстрэп.

2.2. Бустерэп

Метод бутстрэп — это статистический метод для оценки распределения параметров и построения доверительных интервалов. Он основан на идее повторного выбора с возвращением из исходной выборки.

Процедура бутстрэп состоит из следующих шагов:

- Из исходной выборки размером n случайным образом выбираются наблюдения с возвращением, так что новая выборка также имеет размер n.
- На основе этой новой выборки вычисляется интересующая нас статистика (например, среднее значение или медиана).
- Шаги 1-2 повторяются множество раз (обычно от 1000 до 10 000), чтобы получить распределение статистики.
- С помощью полученного распределения статистики строятся доверительные интервалы.

Пример использования метода бутстрэп:
Предположим, у нас есть выборка размером 100 наблюдений и мы хотим построить доверительный интервал для среднего значения. Мы можем применить метод бутстрэп следующим образом:

- Создаем множество бутстрэп-выборок путем случайного выбора 100 наблюдений из исходной выборки с возвращением.
- Для каждой бутстрэп-выборки вычисляем среднее значение.
- Полученные средние значения образуют распределение, которое отражает неопределенность вокруг истинного среднего значения.
- Используем полученное распределение для построения доверительного интервала, например, 95% доверительного интервала будет содержать средние значения, лежащие между 2.5-м и 97.5-м процентилями этого распределения.

Таким образом, метод бутстрэп позволяет оценить статистическую неопределенность и построить доверительные интервалы для различных параметров на основе анализа повторных выборок из исходной выборки.

2.3. Дельта подход

Дельта метод является одним из способов нахождения доверительных интервалов для заданных квантилей. Он основан на аппроксимации нормальным распределением и использует центральную предельную теорему. Аппроксимация нормальным распределением - это метод приближения вероятностного распределения случайной величины нормальным (гауссовским) распределением. Он основан на центральной предельной теореме, которая утверждает, что сумма большого числа независимых и одинаково распределенных случайных величин имеет распределение, близкое к нормальному.

В дельта методе используется аппроксимация первого порядка (линейная аппроксимация) для построения доверительных интервалов. Он предполагает, что функция случайной величины может быть приближена линейной функцией ее математического ожидания и стандартного отклонения. Поэтому доверительный интервал строится, используя оценку математического ожидания и стандартного отклонения случайной величины, взятой с помощью производной этой функции. Дельта метод широко применяется при работе с нелинейными функциями случайных величин и позволяет получить аппроксимацию их доверительных интервалов.

Для нахождения доверительного интервала по дельта методу следуют следующие шаги:

- Вычислить выборочное среднее ($\bar{x}$) и стандартное отклонение (s) на основе имеющейся выборки данных.

- Найти z-значение, соответствующее выбранному уровню доверия, часто это табличные значения (например, для уровня доверия 90% это будет z = 1.645, а для 99% - z = 2.576).

- Рассчитать половину ширины доверительного интервала (δ), умножив z-значение на стандартное отклонение, деленное на квадратный корень из объема выборки: $\delta = z * (s/\sqrt{n})$.

- На основе выборочного среднего и половины ширины интервала можно построить доверительный интервал: $[\bar{x} - \delta, \bar{x} + \delta]$.

Наглядный пример:
Предположим, что у нас есть выборка из 100 наблюдений, и мы хотим построить 90% доверительный интервал для среднего значения. После вычисления выборочного среднего ($\bar{x}$) и стандартного отклонения (s) находим z-значение для 90% доверия, равное 1.645. Затем рассчитываем половину ширины интервала по формуле $\delta = 1.645 * (s/\sqrt{100})$. Итак, если $\bar{x} = 50$ и s = 10, то $\delta = 1.645 * (10/\sqrt{100}) = 1.645$. Таким образом, получаем доверительный интервал [50 - 1.645, 50 + 1.645] = [48.355, 51.645].

Таким образом, дельта метод предоставляет нам инструмент для построения доверительных интервалов для заданных квантилей.

3. Сбор и генерация данных

3.1. Синтезированные данные

При создании тестовых данных для проверки методов поиска доверительного интервала следует учитывать следующие аспекты:

- Известное распределение данных: Выбрать распределение, которое соответствует данным или сценарию исследования. Например, нормальное распределение, равномерное распределение или экспоненциальное распределение. (В нашем случае необходимо проверить все)
- Размер выборки: Определить размер выборки, который будет использоваться для генерации данных. (1000 эл)
- Генерация случайных чисел: Использовать подходящий генератор случайных чисел, чтобы создать данные, соответствующие выбранному распределению и параметрам.
- Повторяемость Убедится, что тестовые данные повторяемы, то есть при каждом запуске теста они будут генерироваться с одинаковыми значениями.

Во время создания программы для генерации данных необходимо учесть, что у каждого метода есть хорошие стороны и обратные, поэтому при синтезе нужно подготовить выборки как одного вида, так и другого. Для этого будем использовать разные распределения: нормальное, равномерное и экспоненциальное. После проверим каждый из методов и проведем анализ. Программа для создания выборок находится в приложении 1 или на GitHub.

3.2. Данные из Kaggle

Также следует проверить методы на данных созданных сообществом программистов и аналитков. Для этого зарегестрируемся на рессурсе Kaggle (Kaggle - это платформа для соревнований по анализу данных и машинному обучению. Она предоставляет сообществу специалистов по данным доступ к наборам данных, инструменты для исследования и моделирования данных, а также возможность участвовать в соревнованиях, где участники решают реальные проблемы с использованием машинного обучения и

статистического анализа. Kaggle также предлагает образовательные материалы, форумы и средства для сотрудничества над проектами.)

4. Анализ и сравнение разобранных методов

4.1. Применение методов

Создадим три программы, каждая из которых на вход получает файл с выборками, находит доверительный интервал, время выболнения метода и False Positive Rate. Выходными данными я вляется фал ".txt"имеющий вид
sample1 90 или 99 доверительный интервал [-1.599138 ; 1.65090] Время поиска 90 интервала 6.91413879395 FPR 0.098 Программы и выходные данные находятся в Приложении 2 или на GitHub.

4.2. Анализ и сравнение методов

Получив результаты от каждого метода сравним их (программа - Приложение 3). Заметим:

4.3. Подведение итогов

5. Заключение

more...

6. Приложение

1. Программа генерации выборок:

DataFactory.py > ...

```python
import numpy as np
import os

# Создание директории для сохранения файлов
save_path = "Users/krivonosovti/Курсач/DataExponential"

# Создание директории для сохранения файлов
if not os.path.exists(save_path):
    os.makedirs(save_path)

num_samples = 150
lambd = 1
# Генерация и сохранение выборок
for i in range(num_samples):

    # Генерация случайной выборки
    #sample = np.random.normal(loc=0, scale=1, size=1000)  # Пример: нормальное распределение
    #sample = np.random.uniform(size=1000)  # Пример: равномерное распределение
    sample = np.random.exponential(scale=1/lambd, size=1000)  # Пример: экспоненциальное распределение
    lambd = lambd / 1.4
    # Создание пути и имени файла
    file_name = f"sampleExponential_{i+1}.txt"
    file_path = os.path.join(save_path, file_name)

    # Сохранение выборки в файл
    with open(file_path, "w") as file:
        for value in sample:
            file.write(f"{value}\n")
```