Санкт-Петербургский государственный университет

Группа 22.М05-мм

Энергетический критерий. Исследование и разработка прикладной библиотеки.

Курлов Дмитрий Николаевич

Отчёт по учебной практике

Научный руководитель:
Профессор кафедры информатики, к. ф.-м. н., В. Б. Мелас

Санкт-Петербург
2023

Оглавление

Введение **3**

1. Постановка задачи **5**

2. Обзор энергетического критерия **6**

2.1. Гипотеза о равенстве двух распределений 6

2.2. Энергетический критерий 6

2.3. Аналитическое исследование предельного распределения 7

2.4. Использование перестановочного метода 9

3. Основные этапы реализации **11**

3.1. Перестановочный метод и эмпирические мощности . . . 11

3.2. Аналитическая формула для теоретических мощностей . 12

4. Результаты моделирования **14**

4.1. Исследовательские вопросы 14

4.2. Метрики 14

4.3. Первые результаты 14

4.4. Обсуждение результатов 16

Заключение **17**

Список литературы **18**

Введение

Современные IT-компании сталкиваются с огромным объемом данных, которые необходимо обрабатывать и анализировать для принятия важных бизнес-решений. DataDriven подход, основанный на анализе данных, является ключевым инструментом для развития успешного бизнеса в настоящее время. Одним из наиболее распространенных методов анализа данных являются A/B-тесты.

A/B-тесты позволяют компаниям проверять гипотезы и принимать решения на основе данных, а не на основе предположений. Они позволяют сравнить две версии продукта или стратегии и определить, какая из них более эффективна. Однако, для проведения A/B-тестов необходимы мощные статистические критерии, которые позволяют определить значимость различий между двумя группами.

В данной работе речь пойдет про исследование и реализацию библиотеки для энергетического критерия, а также расчёта длительности эксперимента.

В работе В. Б. Меласа [5] были определены оптимальные свойства энергетического критерия для проверки гипотезы о равенстве двух распределений, предложенного в работах [1], [3]. Этот критерий применим в случае, когда альтернативное распределение отличается только величиной сдвига. Существует множество тестов, предназначенных для проверки подобных гипотез. Однако, универсальные тесты, такие как тест Колмогорова-Смирнова, могут быть маломощными из-за своей универсальности и показывать слабые результаты для распределений, отличающихся параметром масштаба. Более мощные тесты направлены прицельно на конкретный вид распределений, например, тест Андерсона-Дарлинга.

В данном исследовании были рассмотрены три формулы. Первая формула, после исправления опечаток, является асимптотически точной, что было теоретически доказано в работе научного руководителя. Однако, это обеспечивает ее применимость только в случае больших размеров выборки или незначительных отличий альтернатив. Это подтверждается

результатами вычислительного эксперимента, но при больших отклонениях и размере выборки 100, данная формула значительно (на 10 % и более) превышает эмпирические значения. Однако, с практической точки зрения, ее применимость особенно важна для малых выборок и мощности, близкой к 1. Для этого был введен нормирующий множитель, который представляет собой деление параметра b на $\sqrt{2}$, что эквивалентно умножению дисперсии на данную величину. Это дает практически достаточный эффект, который можно улучшить, с помощью "адаптивного" подбора нормирующего множителя.

1. Постановка задачи

Целью данной работы является исследование энергетического критерия и выведение аналитической формулы для расчёта длительности статистического эксперимента (в дальнейшем формула), а также создание библиотеки обёртки для удобного использования энергетического критерия и формулы. Для достижения цели:

1. Сделать обзор энергетического критерия;
2. Реализовать код для подсчёта эмпирических мощностей;
3. Реализовать расчёт теоретических мощностей;
4. Подобрать нормирующие коэффициенты для аналитической формулы;
5. Создать библиотеку обёртку со следующим функционалом:
    - Моделирование эмпирических мощностей для разных распределений с разными параметрами.
    - Вычисление статистики энергетического критерия, построение её распределения и расчет критических значений для p-value
    - Расчёт длительности эксперимента по заданным параметрам.

2. Обзор энергетического критерия

2.1. Гипотеза о равенстве двух распределений

Рассмотрим классическую задачу проверки гипотезы о равенстве двух распределений

$$H_0 \,:\, F_1 = F_2 \tag{1}$$

против альтернативы

$$H_1 \,:\, F_1 \neq F_2 \tag{2}$$

в случае двух независимых распределений $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ с функциями распределения F_1 и F_2 соответственно [1]. Предположим, что функции распределения F_1 и F_2 принадлежат классу функций распределений со случайной величиной ξ, такой, что

$$E[\ln(1 + \xi^2)] < \infty. \tag{3}$$

Нашей задачей является сравнение мощности различных тестов при проверке гипотезы о равенстве двух распределений с указанными свойствами. Примерами таких распределений являются нормальное распределение, распределение Коши и Лапласа.

Важно отметить, что равенство объемов выборки используется только для удобства обозначений и не влияет на результаты проведения тестирования.

2.2. Энергетический критерий

Рассмотрим следующий тест [4]

$$\Phi_A = \frac{1}{n^2} \sum_{1 \leq i < j \leq n} g(X_i - X_j),$$

$$\Phi_B = \frac{1}{n^2} \sum_{1 \leq i < j \leq n} g(Y_i - Y_j),$$

$$\Phi_{AB} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g(X_i - Y_j),$$

$$\Phi_{nn} = \Phi_{AB} - \Phi_A - \Phi_B, \tag{4}$$

где

$$g(u) = \ln(1 + |u|^2),$$

$g(x)$ с точностью до постоянного члена является логарифмом плотности стандартного распределения Коши.

2.3. Аналитическое исследование предельного распределения

Приведем основные аналитические результаты работы [4]. Рассматривается случай двух распределений, удовлетворяющих свойству (1), отличающихся только параметром масштаба. Для упрощения обозначений предполагается, что $m = n$, а случай $m \neq n$ аналогичен. Критерий (4) принимает вид

$$T_n = \Phi_{nn} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g(X_i - Y_j) - \frac{1}{n^2}\sum_{1\le i<j\le n} g(X_i - X_j) - \tag{5}$$

$$-\frac{1}{n^2}\sum_{1\le i<j\le n} g(Y_i - Y_j) \tag{6}$$

Пусть $f(x)$ обозначает плотность F_1,

$$J(h,n) = \int_R g(x - y - |h|/\sqrt{n}) f(x) f(y) dxdy,$$

$$J_1 = J(0), J_2 = \int_R g^2(x-y) f(x) f(y) dxdy,$$

$$J_3 = \int_R g(x-y) g(x-z) f(x) f(y) f(z) dxdydz,$$

$$J1(h_2, n) = g(x-y)(1 + h_2/n)) f(x) f(y) dxdy.$$

Заметим, что

$$\int_R g'(x-y) f(x) f(y) dxdy = 0, \tag{7}$$

так как функция $g(x)$, по предположению, дважды непрерывно дифференцируема, она может быть представлена в виде $g(x) = \psi(x^2)$, где $\psi(x)$ — дважды непрерывно дифференцируемая функция. Обозначим

$$J^*(h_1) = \frac{1}{2}h_1^2 \int_R g''(x-y)f(x)f(y)dxdy,$$
$$J_1^*(h_2) = \frac{1}{4}h_2^2 \int_R g''_\beta(x - y(1+\beta))|_{\beta=0} f(x)f(y)dxdy.$$

Обозначим

$$b_1^2 = |J^*(h_1)| \tag{8}$$

$$b_2^2 = |J1^*(h_2)| \tag{9}$$

Основным результатом работы [4] является теорема, которая устанавливает вид предельного распределения величины nT_n и представление для асимптотической эффективности теста.

Теорема 1. *Рассмотрим задачу проверки гипотезы* (1)–(2), *где обе функции обладают свойством и симметричны относительно некоторой точки. Тогда*

(i) при условии $n \to \infty$ *функция распределения* nT_n *сходится при* H_0 *к функции распределения случайной величины*

$$(aL)^2 + c \tag{10}$$

где L *- случайная величина, которая имеет стандартное нормальное распределение, где* L *имеет стандартное нормальное распределение, а*

$$a^2 = \sqrt{J_2 + J_1^2 - 2J_3}, c = J_1 - a^2. \tag{11}$$

(ii) Пусть $F_1(x) = F(x), F_2 = F(x(1 + h_2/\sqrt{n}) + h_1/\sqrt{n})$, *где* F — *произвольная функция распределения, симметричная относительно точки и обладающая свойством (1),* h_1, h_2 — *произвольно задан-*

ные числа. Тогда функция распределения nT_n сходится при H_1 к распределению случайной величины

$$(aL + b_1 + b_2)^2 + c,$$

где b_1 имеет вид (8), b_2 определено в (9), а и с заданы формулой (??).

Мощность критерия nT_n с уровнем значимости α асимптотически эквивалентна

$$Pr\{L \geq z_{1-\alpha/2} - b/a\} + Pr\{L \leq -z_{1-\alpha/2} - b/a\}, \qquad (12)$$

где $b = b_1 + b_2$, $z_{1-\alpha/2}$ является таким, что

$$Pr\{L \geq z_{1-\alpha/2}\} = \alpha/2.$$

2.4. Использование перестановочного метода

В исследовании использован перестановочный метод для определения доверительной и критической области значений статистики критерия. Данный метод является статистическим и основан на алгоритме с перестановками внутри выборок.

В рамках данного метода рассматриваются две независимые выборки X и Y, для которых ставится гипотеза о равенстве двух распределений $H_0 : F_1(x) = F_2(x)$ для любого $x \in R$.

Ставится задача нахождения доверительной и критической области значений статистики критерия $\phi = \phi(X, Y)$.

Идея метода заключается в том, что если гипотеза H_0 верна, то элементы выборок X и Y можно рассматривать как элементы из одной выборки, и при перестановке некоторых элементов местами, полученные новые выборки не будут существенно отличаться от исходных.

Алгоритм метода заключается в следующем [2] :

- Объединении выборок X и Y в единую выборку Z;

- Нахождении K случайных разбиений Z на $(X^{(i)}, Y^{(i)})$, таких что: $\#X^{(i)} = \#Y^{(i)}$, $X^{(i)} \cup Y^{(i)} = Z$, $X^{(i)} \cap Y^{(i)} = \varnothing$;

- Нахождении достигнутого уровня значимости теста $ASL_{perm} = \frac{\#\{\phi(X,Y) \leq \phi(X^{(i)},Y^{(i)}), i=1,\ldots,K\}}{K}$;

- Сравнении выбранного уровня значимости α с ASL_{perm}

3. Основные этапы реализации

3.1. Перестановочный метод и эмпирические мощности

В первую очередь, для дальнейшего исследования энергетического критерия и выведение аналитической формулы мощности была написана программа на языке Python, которая реализует подсчёт эмпирической мощности с помощью перестановочного метода. Эта часть нужна для того, чтобы получить достаточно данных для проверки аналитической формулы для мощности, а также для подбора нормирующих коэффициентов.

Основная проблема, которая возникает на данном этапе — это очень долгое время работы перестановочного метода при использовании базовых конструкций языка Python.

Кратко пройдемся по основным этапам работы перестановочного метода для исследования мощности критерия:

1. Генерация выборок с заданными параметрами: размер, математическое ожидание, дисперсия;
2. Подсчёт "эталонной" статистики критерия для сгенерированных выборок;
3. Объединение выборок в единую;
4. Получение K пар случайных подвыборок из большой результирующей выборки;
5. Подсчёт статистики критерия для каждой K-й пары и сравнение с эталонной;
6. Сравнение отношения "расходящихся случаев" с 0.05;

Если для получения случайных выборок существуют уже исследованные методы — специализированные библиотеки NumPy и SciPy, то для подсчёта статистики энергетического критерия таких нет. В

свою очередь, статистику необходимо подсчитывать $N = KNn_h$ раз, где K – количество разбиений на каждой итерации, N – количество экспериментов в моделировании, n_h - количество шагов в смещении. Для получения достоверных результатов эмпирически было подобраны следующие параметры: $K = 700$, $N = 1000$, $n_h = 10$

Для улучшения производительности программы были выбраны следующие инструменты:

1. Использование векторных операций и массивов NumPy
2. Использование параллельных вычислений

Стратегий распараллеливания в данном случае достаточно много. В базовой реализации использовалось распараллеливание по шагам смещения.

3.2. Аналитическая формула для теоретических мощностей

Одной из важных компонент библиотеки будет являться аналитическая формула для расчёта длительности эксперимента.

Для расчета длительности эксперимента используют формулы MDE (Minimum Detectable Effect), которые учитывают размер выборки, уровень значимости, мощность теста и ожидаемый эффект. Определение MDE позволяет определить минимальный размер выборки и длительность эксперимента для достижения статистической значимости и получения надежных результатов.

Для расчета длительности эксперимента с помощью MDE необходимо выполнить следующие шаги:

1. Определить ожидаемый эффект (Expected Effect Size), то есть насколько большим должно быть различие между контрольной и тестовой группами, чтобы результаты эксперимента были значимыми;

2. Выбрать уровень значимости (Significance Level), который определяет вероятность ошибки первого рода, то есть отвержения нулевой гипотезы, когда она на самом деле верна.;

3. Выбрать мощность теста (Power), которая определяет вероятность обнаружения эффекта, когда он действительно существует. Обычно используют мощность теста 0,8 или 0,9;

4. Определить размер выборки (Sample Size), который необходим для достижения заданного уровня значимости и мощности теста при заданном ожидаемом эффекте;

Сама длительность в днях рассчитывается на основе полученного Sample Size. Sample Size делится на поток наблюдений за единицу времени. В каждой конкретной компании и каждом случае поток наблюдений индивидуальный и заранее известный.

Для получения формулы MDE в случае энергетического критерия можно воспользоваться теоремой 1. В рамках теоремы нужно расчитать ряд коэффициентов, которые в общем случае являются интегралами и зависят от размера сравниваемых выборок и величины "смещения". Для вычисления этих интегралов воспользуемся методом Монте-Карло.

4. Результаты моделирования

4.1. Исследовательские вопросы

Для успешного завершения работы необходимо следующее:

- Написать код, который реализует подсчёт эмпирических и теоретических мощностей энергетического критерия. Критерий должен рассчитываться достаточно быстро;
- Эмпирические мощности должны биться с теоретически рассчитанными, что будет гарантировать корректность формулы

4.2. Метрики

В данной работе наиболее важна скорость выполнения и точность аналитической формулы. Поэтому для понимания успешности будут использоваться следующие метрики:

- Время моделирования эмпирических и теоретических мощностей;
- Среднее абсолютное отклонение (Mean Absolute Error) между теоретическими и эмпирическими мощностями;
- Сравнение эмпирических мощностей для случаев различных распределений и различных статистических критериев;

4.3. Первые результаты

Ниже представлены результаты моделирования. Получены первые эмпирические и теоретические мощности для распределений с тяжелыми хвостами для случая сдвига, так как случай сдвига является наиболее важным для индустрии. Также посчитаны первые результаты времени выполнения программы.

Таблица 1: Различия между теоретическими и эмпирическими мощностями. Число итераций $N = 1000$, объем выборок $n = 100$, число перестановок $K = 700$, уровень значимости $\alpha = 0.05$. Для моделирования соответствующей $H1$ ситуации использовались распределения $Cauchy(0,1)$ и $Cauchy(h/\sqrt{n}, 1)$

h	Эмп. мощность	Теор. мощность	Отклонение
0	0.05	0.049	0.01
1	0.074	0.055	0.019
2	0.122	0.072	0.05
3	0.223	0.124	0.099
4	0.400	0.244	0.156
5	0.554	0.300	0.254
6	0.708	0.455	0.253
7	0.819	0.602	0.217
8	0.898	0.709	0.189
9	0.960	0.800	0.16
10	0.970	0.878	0.092

Таблица 2: Сравнение скорости выполнения с использованием базовых средств языка VS с использованием векторных операций. Выборка из 100 наблюдений. Параметры выборки $h = 5$, $n = 100$, число перестановок $K = 50$

Базовая реалзиация	Векторные операции	Разница
2.71 $\pm$0.06 мин	1.51 $\pm$0.01 сек	100+ раз

4.4. Обсуждение результатов

По результатам на текущий момент. Удалось построить оптимальную реализацию подсчёта критерия и перестановочного метода, которая позволяет за обозримое время проводить достоверное моделирование мощностей.

Реализована первая версия подсчёта теоретической мощности. Формула нуждается в дальнейшей доработке и подборе нормирующих параметров.

Заключение

В данной работе рассмотрена предметная область и поставлена задача. Приведён необходимый теоретический обзор энергетического критерия. Реализован подсчёт эмпирических и теоретических мощностей. Получены первые результаты моделирования. Удалось построить достаточно оптимальную реализацию расчёта эмпирического эксперимента. План дальнейшего продвижения

- Доработка формулы для теоретической мощности и расчёта длительности эксперимента.
- Начать реализацию библиотеки обёртки.

Список литературы

[1] Aslan B., Zech G. New test for the multivariate two-sample problem based on the concept of minimum energy // Journal of Statistical Computation and Simulation. — 2005. — Vol. 75, no. 2. — P. 109–119.

[2] Efron B., Tibshirani R.J. An Introduction to the Bootstrap. — Springer Science+Business Media Dordrecht, 1993. — P. 202–218.

[3] Melas V. Salnikov D. On Asymptotic Power of the New Test for Equality of Two Distributions // Recent Developments in Stochastic Methods and Applications. — 2021. — Vol. 371. — P. 204–214.

[4] Мелас В.Б. Об асимптотической мощности одного метода проверки гипотез о равенстве распределений // Вестник СПбГУ. — 2022. — Т. 4.

[5] Мелас В.Б. Об оптимальных свойствах энергетического критерия проверки гипотез о равенстве распределений // Вестник СПбГУ, Сер. 1, Вып. 2. (в печати). — 2023.