Санкт-Петербургский государственный университет
Математическое обеспечение и админимтрирование информационных систем

Учебная практика 1 (научно-исследовательская работа)

Численная проверка применимости формулы для асимптотической мощности

Выполнил:
Курлов Дмитрий Николаевич
группа 22.М04-мм

Научный руководитель:
д. ф.-м. н., профессор
Мелас Вячеслав Борисович
Кафедра Статистического Моделирования

Санкт-Петербург
2023

Оглавление

Введение

В работе В. Б. Меласа [1] была получена формула для теоретической мощности энергетического критерия для проверки гипотезы о равенстве двух распределений, предложенного в работах [2], [3]. Этот критерий применим в случае, когда альтернативное распределение отличается только величиной сдвига.

В работе В. Б. Меласа [4] рассматривается уже случай изменения параметра масштаба. Также была выведена соответствующая формула для теоретических мощностей. 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.

Данная работа заключается в исследовании случая изменения параметра масштаба. Необходимо провести численное моделирование эмпирических мощностей для различных значений параметра масштаба, а затем сравнить результаты с полученной теоретической формулой.

Глава 1

Постановка задачи

1.1. Гипотеза о равенстве двух распределений

Рассмотрим классическую задачу проверки гипотезы о равенстве двух распределений

$$H_0 : F_1 = F_2 \tag{1.1}$$

против альтернативы

$$H_1 : F_1 \neq F_2 \tag{1.2}$$

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 [2]. Предположим, что функции распределения F_1 и F_2 принадлежат классу функций распределений со случайной величиной ξ, такой, что

$$E[\ln(1 + \xi^2)] < \infty. \tag{1.3}$$

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.

1.2. Энергетический критерий

Рассмотрим следующий тест [1]

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

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

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

$$\Phi_{nn} = \Phi_{AB} - \Phi_A - \Phi_B, \tag{1.4}$$

где

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

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

1.3. Аналитическое исследование предельного распределения

Приведем основные аналитические результаты работы [1]. Рассматривается случай двух распределений, удовлетворяющих свойству (1.3), отличающихся только параметром масштаба. Для упрощения обозначений предполагается, что $m = n$, а случай $m \neq n$ аналогичен. Критерий (1.4) принимает вид

$$T_n = \Phi_{nn} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g(X_i - Y_j) - \frac{1}{n^2}\sum_{1\le i<j\le n} g(X_i - X_j) - \tag{1.5}$$

$$-\frac{1}{n^2}\sum_{1\le i<j\le n} g(Y_i - Y_j) \tag{1.6}$$

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

$$\int_R g'(x - y) f(x) f(y) dxdy = 0, \tag{1.7}$$

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

$$b_1^2 = |J^*(h_1)| \tag{1.8}$$

$$b_2^2 = |J1^*(h_2)| \tag{1.9}$$

Основным результатом работы [1] 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 (3) и имеют плотности распределения симметричные относительно некоторой точки. Тогда*

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

$$(aL)^2 + c, \tag{1.10}$$

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

$$c = J_1 - a^2, a^2 = \sqrt{J_2 + J_1^2 - 2J_3}, \tag{1.11}$$

(ii) Пусть $F_1(x) = F(x), F_2(x) = F(x(1 + h_2/\sqrt{n}) + h_1/\sqrt{n})$, где F — произвольная функция распределения,с плотностью $f(x)$ симметричной относительно некоторой точки и обладающая свойством (3), h_1, h_2 - произвольные заданные числа .

Тогда функция распределения nT_n сходится при выполнении гипотезы H_1 к распределению случайной величины

$$(aL + b_1 + b_2)^2 + r(h_1, h_2)L + c, \tag{1.12}$$

*где b_1 имеет вид (**??**), b_2 определено в (**??**), a и c заданы формулой (1.11), $r(h_1, h_2)L$ - остаточный член, который имеет громоздкое выражение, но предположительно мало влияет на мощность теста.*

Мощность критерия nT_n с уровнем значимости α при отбрасывании остаточного члена $r(h_1, h_2)L$ асимптотически эквивалентна

$$Pr\{L \geq z_{1-\alpha/2} - b/a\} + Pr\{L \leq -z_{1-\alpha/2} - b/a\}, \tag{1.13}$$

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

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

Глава 2

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 $\varphi = \varphi(X, Y)$.

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 [5] :

- 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 $ASL_{perm} = \frac{\#\{\varphi(X,Y) \leq \varphi(X^{(i)},Y^{(i)}), i=1,\ldots,K\}}{K}$;
- Сравнении выбранного уровня значимости α с ASL_{perm}

Заключение

В данной работе была исследована теоретическая часть, необходимая для моделирования эмпирических мощностей. А также аналитические результаты теоретических работ, в которых были получены асимптотические формулы для энергетического критерия.

В данный момент ведется разработка ПО для проведения экспериментов и моделирования эмпирических мощностей энергетического критерия для случая изменения параметра масштаба для различных распределений. Также данное ПО будет расчитывать теоретические мощности и сравнение результатов.

Список литературы

1. Мелас В.Б. Об асимптотической мощности одного метода проверки гипотез о равенстве распределений. // Вестник СПбГУ, Вып. 2 (в печати). — 2023.
2. 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.
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 „энергетического“ теста для проверки гипотез о равенстве распределений // Вестник СПбГУ, Сер. 1, Вып. 2. (в печати). — 2023.
5. Efron B., Tibshirani R.J. An Introduction to the Bootstrap. — Springer Science+Business Media Dordrecht, 1993. — P. 202–218.